UNIVERSIDAD AUTONOMA DE MADRID

Department of Biochemistry

# Epigenetic and Transcriptional Variability in Leukemia and Normal Blood Cells: A Computational Exploration

Simone Ecker

Madrid, 2015


Department of Biochemistry
Faculty of Medicine
Autonomous University of Madrid

# Epigenetic and Transcriptional Variability in Leukemia and Normal Blood Cells: A Computational Exploration

A Thesis Submitted in
Partial Fulfillment of the Requirements for the
Degree of

**"PhD in Biochemistry, Molecular Biology, Biomedicine and Biotechnology (Molecular Biosciences)"**

by
Simone Ecker
Diplom-Ingenieur of Biomedical Informatics

Thesis directors:
Univ.-Prof. Dr. Alfonso Valencia Herrera
Dr. Daniel Rico Rodriguez

Spanish National Cancer Research Center (CNIO)
Madrid, 2015

## Confirmation of the Supervisor

I hereby declare to have supervised the present thesis and consequently approve its submission with a positive assessment.

.....

Date and signature of the first supervisor

.....

Name of the first supervisor in upper-case letters

.....

Date and signature of the second supervisor

.....

Name of the second supervisor in upper-case letters

*"The human genome, like the universe, is composed of only a tiny fraction of readily understood regions. The rest – more than 98 percent – is largely shrouded in mystery. Just as physicists probe the universe in search of dark matter and dark energy, epigeneticists are exploring the parts of the genome outside of our genes for various pieces of evidence that can shed light on new corners."*

— Haley Bridger, *Broad Communications*

# Acknowledgement

The completion of this thesis, and all the work behind it, would not have been possible without the support of many people. First and foremost, I want to express my sincere gratitude to my thesis director Univ.-Prof. Dr. Alfonso Valencia for giving me the opportunity to conduct this research, for encouraging comments and discussions, and for always being there when any kind of help was needed, often even in the middle of the night, especially during critical phases in the preparation of presentations or manuscripts. I am very grateful for the continuous support and helpful advice that he has provided.

Also to Dr. Daniel Rico, co-director of the present thesis and staff scientist in our research group, I would like to express my thankfulness for offering his guidance and advice on the research projects carried out during my PhD studies, the insightful comments and discussions on the biology behind, his great ideas, and for all his patience and support throughout the years.

Very special thanks go to Dr. Vera Pancaldi, postdoctoral researcher in our team, in particular for the great collaboration on the CLL work, for all the useful discussions, and for always answering my questions and listening to me, even during her maternity leave. It has been a real pleasure working together with her, and I am very grateful for all I could learn due to her incredible expertise in the field of biological noise and variability.

I also want to thank all my other colleagues in our research group, as well as my PhD thesis committee – Dr. Ana Losada, Dr. Ramón Díaz-Uriarte, and Dr. Iñaki Martín-Subero – for useful discussions and advice on research and my dissertation. Additional thanks go to Dr. Iñaki Martín-Subero for the very interesting and fruitful collaboration on CLL and B cell research in the context of the ICGC and BLUEPRINT.

Furthermore, I express my gratitude to Univ.-Prof. Dr. Stephan Beck for giving me the opportunity to work with him and his research group at University College London Cancer Institute on BLUEPRINT. My stay there was a fantastic experience from which I gained a lot of inspiration and motivation for the last part of my PhD studies. In particular, I thank Dr. Dirk Paul for the outstanding collaboration on the project, and all the members of the team for making my stay in their research group such a pleasure.

I am also very thankful for the support of all the members of the BLUEPRINT consortium, especially those involved in work package 10, and all other people who have contributed to my work in one way or another.

Beside all research collaborators, I would like to thank the Bioinformatics Unit of the CNIO, and especially José María Fernández, bioinformatics technician at the National Bioinformatics Institute, for their endless support in administrative questions regarding the computer systems at the institute, for finding out about failing Unix libraries causing R to forget how basic algebra works, and for spending an uncountable amount of hours in providing support to resolve problems with system failures, magic bash scripts, and broken hard disks.

For financial support I want to thank the "La Caixa" Foundation from which I have received an International PhD Programme Fellowship to realize this work.

Last but not least, I express my very special thanks to all my friends for their continuous support and unconditional friendship despite my constant lack of time because of work, or thesis writing. I am incredibly grateful for their understanding and all their patience. Particularly, I also want to deeply thank my parents for supporting me throughout my life in whatever adventure I start.

**Resumen**

En todos los sistemas biológicos están implicados procesos estocásticos. Hoy en día, se sabe que la variabilidad biológica es necesaria para controlar el comportamiento de un sistema multicelular en su conjunto, y para permitir la adaptación rápida a cambios en el ambiente. Se ha demostrado que la plasticidad fenotípica es clave en el funcionamiento del sistema inmunológico, pero que también está fuertemente asociada con enfermedades, especialmente cáncer. Hasta el momento, se han investigado principalmente factores genéticos en este contexto, y sólo pocos estudios se han centrado en la heterogeneidad a nivel epigenómico y transcriptómico.

En el marco de dos consorcios internacionales, utilizamos los primeros conjuntos de datos de gran escala disponibles para explorar la variabilidad biológica. Desarrollamos nuevas estrategias analíticas que combinan diferentes métodos para medir la variabilidad con modelos estadísticos bien establecidos que nos permiten conseguir una cuantificación robusta de la heterogeneidad interindividual. Estudiamos la variabilidad en dos contextos biológicos distintos. Primero, cuantificamos la variabilidad transcripcional en grandes cohortes de pacientes con leucemia linfocítica crónica (LLC) y estudiamos las diferencias entre los dos subtipos principales de la enfermedad. Segundo, analizamos la variabilidad en la metilación del ADN y la expresión génica en monocitos, neutrófilos y células T obtenidos de 48 individuos sanos.

Encontramos que el subtipo más agresivo de LLC muestra un aumento significativo de la heterogeneidad transcripcional interindividual. Los genes con una mayor variabilidad en la forma agresiva de la enfermedad están enriquecidos en funciones relacionadas con el ciclo celular, rutas de señalización, diferenciación celular, y desarrollo. Estas observaciones indican una posible relación entre la heterogeneidad transcripcional y la progresión y agresividad de la enfermedad. El análisis de la variabilidad diferencial entre los tres tipos celulares más abundantes del sistema inmunológico humano mostró que los neutrófilos tienen mayor variabilidad tanto en sus perfiles de metilación del ADN como en los de expresión génica. Esto puede deberse a la función de los neutrófilos como las primeras células del sistema inmune que migran a sitios de inflamación, ya que la plasticidad epigenética y transcripcional son esenciales para permitir una respuesta rápida a cambios en el entorno.

En conjunto, los resultados de este trabajo recalcan la importancia de la variabilidad epigenética y transcripcional en el sistema inmunológico humano tanto en condiciones sanas como enfermas, así como la necesidad del desarrollo de técnicas bien fundamentadas para analizar dicha variabilidad. Nuestros resultados proporcionan nuevos conocimientos sobre la plasticidad de las células del sistema inmune normales y con LLC, y a largo plazo, el estudio de la heterogeneidad a nivel epigenómico y transcriptómico habilitará el desarrollo de estrategias terapéuticas dirigidas a modular la variabilidad en enfermedades hematopoyéticas e inmunológicas.

**Abstract**

Stochastic processes are involved in every biological system. Nowadays it is well known that biological variability is necessary to reliably control the behavior of a multicellular system as a whole, and to enable a rapid adaptation to changes in the environment. Phenotypic plasticity has been demonstrated to be key in the functioning of the human immune system, but has also been strongly associated with human diseases, especially cancer. Until now, mainly genetic factors have been investigated in this context, and only few studies focused on heterogeneity at the epigenomic and transcriptomic level.

Within the framework of two international genome consortia, we used the first available large-scale datasets to explore biological variability. To this aim, we developed new analytical approaches combining different methods to measure variability with well-established statistical models to achieve a robust quantification of interindividual heterogeneity. We studied variability at different levels in two distinct biological contexts. First, we quantified interindividual gene expression variation in two large cohorts of chronic lymphocytic leukemia (CLL) patients and studied differences between the two main subtypes of the disease. Second, we analyzed DNA methylation and gene expression variability across primary human monocytes, neutrophils, and T cells derived from 48 healthy individuals.

We found that the more aggressive subtype of CLL shows significantly increased gene expression heterogeneity across patients. Genes with increased variability in the aggressive form of the disease are strongly enriched in functions related to the cell cycle and show furthermore significant associations with signaling, cell differentiation, and development. These observations indicate a possible relation between heterogeneous gene expression patterns and disease progression and aggressiveness. Analyzing differential variability across the three most abundant cell types of the human immune system, we found that neutrophils show increased variability in both their DNA methylation and gene expression patterns. We hypothesized that this is due to the neutrophils' function as the first responders in the immune system that migrate to sites of inflammation, as epigenetic and transcriptional plasticity are essential to enable rapid adaptation to new and changing environments.

Taken together, the results of this work highlight the importance of epigenetic and transcriptional variability in the human immune system in health and disease, as well as the necessity of the development of well-founded techniques to analyze this variability. Our findings provide new insights into the plasticity of CLL and normal immune cells, and in the long run the study of heterogeneity at the epigenomic and transcriptomic level will empower the development of therapeutic strategies aiming to modulate variability in hematopoietic and immunological diseases.

# Contents

**Abbreviations** **3**

**1 Introduction** **5**

1.1 Gene Expression ..... 5
1.2 Epigenetic Modifications ..... 7
1.3 Epigenetic and Transcriptional Variability ..... 11
1.3.1 Introduction to Biological Variability ..... 11
1.3.2 Biological Significance and Functions of Variability ..... 12
1.3.3 Variability in Cancer ..... 15
1.4 Human Blood Cells ..... 17
1.5 Chronic Lymphocytic Leukemia ..... 19
1.5.1 The Disease of Chronic Lymphocytic Leukemia ..... 19
1.5.2 Variability in Chronic Lymphocytic Leukemia ..... 23

**2 Objectives** **25**

**3 Methods** **27**

3.1 Gene Expression Variability in CLL ..... 27
3.1.1 Material ..... 27
3.1.2 Measuring Gene Expression Variability ..... 28
3.1.3 Analysis of DNA Methylation and its Relationship to Gene Expression Variability ..... 29
3.1.4 Functional Analysis ..... 30
3.1.5 Random Forest Classification ..... 30
3.1.6 Programming Language ..... 30
3.2 Variability in Normal Blood Cells ..... 31
3.2.1 Material ..... 31
3.2.2 Data Preprocessing and Filtering ..... 32

3.2.3 Measuring DNA Methylation and Gene Expression ..... 33
3.2.4 Analysis of DNA Methylation Variability ..... 34
3.2.5 Analysis of Gene Expression Variability ..... 34
3.2.6 Analysis of Sex-Specific Differential Expression and DNA Methylation 35
3.2.7 Programming Language ..... 36

**4 Results & Discussion** **37**
4.1 Gene Expression Variability in CLL ..... 37
4.1.1 Overview ..... 37
4.1.2 Measuring Gene Expression Variability ..... 38
4.1.3 Gene Expression Variability in the Two Subtypes of CLL ..... 40
4.1.4 Gene Expression Variability and DNA Methylation ..... 42
4.1.5 Functional Analysis of Differentially Variable Genes ..... 44
4.1.6 Classification of Patients by Gene Expression Variability ..... 52
4.1.7 Interpretation and Further Discussion ..... 57
4.2 Variability in Normal Blood Cells ..... 59
4.2.1 Overview ..... 59
4.2.2 Analysis of Variability in Different Biological Data Types ..... 59
4.2.3 Comparison of Statistical Methods to Analyze Differential DNA Methylation Variability ..... 61
4.2.4 DNA Methylation Variability in Normal Blood Cells ..... 64
4.2.5 Gene Expression Variability in Normal Blood Cells ..... 72
4.2.6 Sex-Specific Differential Expression in Normal Blood Cells ..... 76
4.2.7 Relationship Between DNA Methylation Variability and Gene Expression ..... 81
4.2.8 Interpretation and Further Discussion ..... 85
4.2.9 Outlook ..... 87

**5 Conclusions** **90**

**List of Figures** **94**

**List of Tables** **97**

**Bibliography** **98**

**Annex I** **140**

**Annex II** **176**

# Abbreviations

**AML** acute myeloid leukemia
**AUC** area under the curve
**BCR** B cell receptor
**CLL** chronic lymphocytic leukemia
**CNV** copy number variation
**CV** coefficient of variation
**DE** differentially expressed
**DNA** deoxyribonucleic acid
**DNMT** DNA methyltransferase
**DV** differentially variable
**EV** expression variability measured by the method of Alemu *et al.* (2014)
**FDR** false discovery rate
**GC** germinal center
**GEO** gene expression omnibus
**GO** gene ontology
**HIV** human immunodeficiency virus
**ICGC** International Cancer Genome Consortium
**$\mathbf{IgV_H}$** immunoglobulin variable-region heavy chain
**IHEC** International Human Epigenome Consortium
**IQR** interquartile range
**M-CLL** $IgV_H$ "mutated" CLL
**MAD** mean absolute deviation
**MHC** major histocompatibility complex
**MMTV** mouse mammary tumor virus
**MV** methylation variability measured by the method of Alemu *et al.* (2014)
**NET** neutrophil extracellular trap
**NK** natural killer
**NPC** nuclear pore complex
**PBMC** peripheral blood mononuclear cells

**RMA** robust multi-array average
**RNA** ribonucleic acid
**RNAseq** RNA sequencing
**RRBS** reduced representation bisulfite sequencing
**SD** standard deviation
**SNP** single nucleotide polymorphism
**SWAN** subset-quantile within array normalization
**TCR** T cell receptor
**TRAIL** tumor necrosis factor related apoptosis-inducing ligand
**U-CLL** $IgV_H$ "umutated" CLL
**UTR** untranslated region
**WGBS** whole genome bisulfite sequencing
**WGS** whole genome sequencing

# Chapter 1

# Introduction

## 1.1 Gene Expression

Every cell in a multicellular organism contains an identical copy of the genetic blueprint. Nevertheless, cells differ dramatically in terms of shape, size and function. Different cell types can develop because the cells synthesize different RNA and therefore different protein molecules (Alberts *et al.*, 2004). This is what we call gene expression, part of the central dogma of molecular biology: "DNA makes RNA makes protein" (Crick, 1970).

More exactly, gene expression is the process by which the genetic information at the level of a genes' DNA sequence is transcribed into RNA to subsequently produce a functional gene product such as a protein. It is the most fundamental level at which the genotype (that is, the genetic makeup of a cell) gives rise to the phenotype (the set of an organism's observable characteristics). Thus, different types of cells may possess different gene expression profiles although they all have the same genomic sequence (Alberts *et al.*, 2004).

This is possible due to gene regulation, forming the basic principle for cellular differentiation, and for the versatility and adaptability of any organism. The cell can activate (or upregulate) or repress (or downregulate) genes in response to a phase of the cell cycle, a developmental stage, or to adapt to the environment and external signals such as temperature changes or the treatment with a hormone (Alberts *et al.*, 2004; Bird, 2007; Reik, 2007).

Every single cell adjusts the speed and rate of the expression of different genes according to its needs (see figure 1.1 for an illustrative example). The amount and timing of the production of genes is influenced by different mechanisms that control the transcription of DNA into RNA, described in Alberts *et al.* (2004), as well as Zaidi *et al.* (2004); Mattick *et al.* (2009); Martinez & Walhout (2009), and summarized in the following two paragraphs.

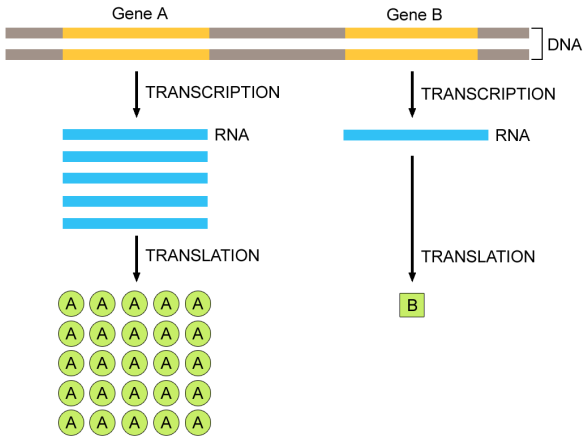


Figure 1.1: Genes can be expressed with different efficiencies. Gene A gets transcribed and translated at much higher rates than gene B, which allows the amount of protein A to be much greater than that of protein B.

Figure adapted from Alberts *et al.* (2004).

The production of RNA is performed by the enzyme RNA polymerase which reads the DNA to transcribe it into complementary RNA. Sections of the DNA can be more or less accessible or even hidden from the transcriptional machinery due to its structure, and the folding and packaging of the DNA. These local characteristics of the DNA and consequently also the gene expression activity in these regions are furthermore heavily influenced by epigenetic modifications, described in section 1.2. Additionally, there are proteins that can attach themselves dynamically at regions of the DNA and thus influence their transcription. Those proteins are called transcription factors. They can suppress or increase the activity of a gene.

Moreover, gene transcription gets regulated through additional specificity factors, repressors, activators and enhancers, which can further control when transcription occurs and how much RNA is produced. The amount of RNA that gets finally translated into protein is determined by translation, the process of the synthesis of RNA into proteins. This process is regulated in a similar way as transcription, for example via translational repressors that may inhibit the initiation of translation due to blocking the starting point.

All the above described principles of gene regulation can interact among themselves and thereby build highly complex regulating mechanisms. The examination of gene expression patterns on the genomic scale helps to understand the mechanisms that control multicellular development and pathological cellular events, and can provide important clues to gene function in health and disease (Schena, 2003; Rockman & Kruglyak, 2006).

## 1.2 Epigenetic Modifications

The concept of epigenetics was first introduced by Waddington (1939, 1942) as "the causal interactions between genes and their products, which bring the phenotype into being". Later, the term epigenetics was re-defined as heritable changes in gene expression that are not caused by alterations in the underlying DNA sequence (Holliday, 1987). The word "epi" is of Greek origin ("επί") and means "over", "outside", or "on top of". Thus, epigenetic modifications are modifications on top of the genetic material. These modifications mark the genome using chemical compounds, which leads to alterations in the transcriptional potential of a cell. They may last through cell divisions for the duration of a cell's life, and they may also be inherited to subsequent generations (Bird, 2007).

Higher multicellular organisms, such as mammals or many plants, have large genomes in which as much as half of the genes can be transcriptionally deactivated in particular cell types (Cedar & Bergman, 2011). To achieve this, and to tightly control cellular processes, a complex regulatory strategy is necessary (Cedar & Bergman, 2011). In contrast with the genome, the epigenome is highly dynamic. All cells of a multicellular organism are characterized by essentially the same genome, but many different epigenomes, which influence which genes are active and therefore also which proteins can be produced in every particular cell (Adams *et al.*, 2012).

Nowadays, it is well known that epigenetic modifications are key in cellular differentiation and development (Zhu *et al.*, 2013; Jones, 2012), that the epigenome changes throughout lifetime (Horvath, 2013; Heyn *et al.*, 2012; Hannum *et al.*, 2013), and that it serves as the intersection between the genome and the environment (Bonasio *et al.*, 2010; Lam *et al.*, 2012). Food, lifestyle changes and psychological factors such as stress have been shown to be able to alter the epigenome, and in some studies these changes have even been shown to be passed on to subsequent generations (Dias & Ressler, 2014; Veenendaal *et al.*, 2013; Morgan *et al.*, 1999; Lam *et al.*, 2012).

It is also very well established nowadays that epigenetic modifications are a hallmark of cancer (Sharma *et al.*, 2009; Timp & Feinberg, 2013; Esteller, 2008). As epigenetic modifications are a dynamic and furthermore reversible process, they are a promising target for therapeutic approaches (Witte *et al.*, 2014; Baylin & Jones, 2011).

The so far best characterized epigenetic modifications are histone modifications and DNA methylation, which will be described in more detail in the next few paragraphs, starting with an introduction to histone modifications.

Genomic DNA is packaged into complexes of protein and DNA, called chromatin. The prevalent type of protein found in chromatin are histones, which condense the DNA. The primary units of the chromatin structure, the nucleosomes, consist of an octamer of histone proteins (H2A, H2B, H3 and H4), where the DNA is wrapped around (Lay *et al.*, 2015; Barski *et al.*, 2007). The nucleosomes and the DNA form a chromatin fiber which can be further condensed (Barski *et al.*, 2007; Kouzarides, 2007). The organization of the chromatin is important to maintain the balance between compaction and accessibility of the genome (Lay *et al.*, 2015; Li *et al.*, 2007).

Chromatin packaging of DNA varies depending on the cell cycle stage and by local DNA region (Lay *et al.*, 2015). Most DNA is packaged in a closed, tightly packed chromatin conformation, so-called heterochromatin (Kouzarides, 2007). Actively transcribed genes however have to be highly accessible to transcription factors and other DNA binding proteins. Such unpackaged or loose chromatin is called euchromatin and usually transcriptionally active (Kouzarides, 2007). Thus, the chromatin structure influences gene expression by making genomic regions more or less accessible for the transcription machinery of the cell via remodeling the chromatin structure and changing the density of packaging (Kouzarides, 2007; Cedar & Bergman, 2011).

This chromatin remodeling occurs via post-translational modifications of the long amino acid chains that make up the histone proteins. The to date most highly studied modification of histone tails is acetylation, but many additional modifications are known, for example histone methylation, phosphorylation, ubiquitination and SUMOylation, among others (Helin & Dhanak, 2013).

Different histone modifications have distinct regulatory functions. They can influence for example transcriptional initiation or elongation, enhancer activity, or transcriptional repression (Ernst *et al.*, 2011). Furthermore, the genomic context where the histone modification occurs can lead to different effects. For example, the same modification may be able to activate a gene when lying in the coding section of its body, while acting as a transcriptional inhibitor when found at the gene's promoter region (Kouzarides, 2007).

The complete set of histone modifications in a cell is known as the histone code (Strahl & Allis, 2000), and the combination of the modifications can provide a more precise insight into chromatin states and their regulatory functions (Jenuwein & Allis, 2001; Ernst & Kellis, 2010).

DNA methylation is the most widely studied and well-characterized epigenetic modification to date (Lam *et al.*, 2012; Jones, 2012). It is the process by which a methyl group is added to the DNA by DNA methyltransferases (DNMTs), which can be reversed by an antagonistic group of enzymes, DNA de-methylases (Jones, 2012). The most common type of methylation is the conversion of cytosine into 5-methylcytosine (5mC) at CpG dinucleotides, that is, cytosine followed by guanine in the DNA sequence, with only a phosphate in between (Laird, 2010).

Some areas of the genome are more heavily methylated than others, and CpGs are not evenly distributed across the genome (Laird, 2010). Regions with a high frequency of CpG sites are called CpG islands (Jones, 2012). One of the first observations that have been made studying DNA methylation was that when it occurs at CpG islands in the promoter region of a gene, it has the effect of repressing gene expression, in contrast to unmethylated promoter regions, which are associated to active transcription (Laird, 2010; Riggs, 1975; Holliday & Pugh, 1975), see figure 1.2.

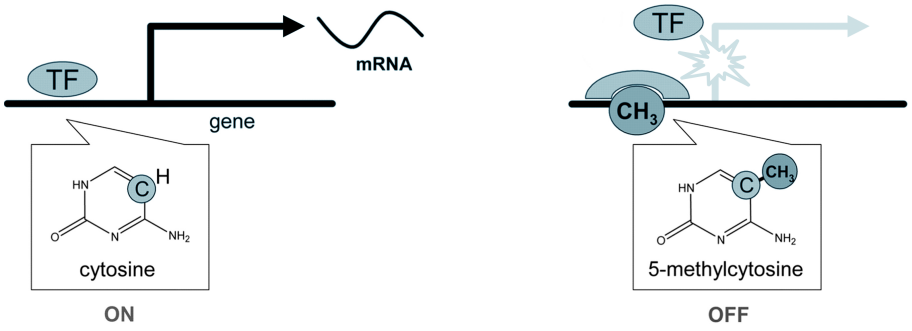


Figure 1.2: DNA methylation at a gene's promoter can silence its expression. The methylation of the cytosine attracts capping proteins that hinder access for transcription factors that normally turn on gene expression. When the transcription factor cannot bind to the promoter area of the gene, transcription of RNA does not occur, and the gene is silenced.
Figure adapted from Zeisel (2007).

This is best understood for tumor suppressor genes in cancer (Esteller, 2008). However, with the establishment of high-throughput technologies to investigate DNA methylation profiles genome-wide, the traditional view of DNA methylation being a silencing epigenetic mark has been challenged (Jones, 2012). The relationship between DNA promoter methylation and gene expression has shown to be less straightforward in non-malignant tissues, and generally, the genome-wide correlation of gene promoter methylation and gene expression levels is often very low (Lam *et al.*, 2012).

Another question that has not yet been fully elucidated is the one of "what comes first?", DNA methylation that causes gene silencing, or the deactivation of a gene followed by the process of methylation to add an additional layer of stability to the silencing. In several studies, DNA methylation indeed appeared to serve as a "lock" to reinforce the state of already inactive genes, demonstrated for example in X chromosome inactivation and tumor suppressor genes (Lock *et al.*, 1987; Clark & Melki, 2002; Widschwendter *et al.*, 2007; Ohm *et al.*, 2007; Schlesinger *et al.*, 2007; Jones, 2012). Others have shown that DNA methylation can also have a more instructive role, for example in the initiation of silencing in hematopoietic stem cell differentiation (Byun *et al.*, 2009; Jones, 2012).

Traditionally, most DNA methylation studies focused on methylation occurring at CpG islands associated to gene promoter regions, but now it has become apparent that methylation located in the coding region of a body of a gene, or in intergenic regions containing enhancers and insulators, all of which are typically present in CpG depleted regions, exhibits crucial functions in development, differentiation and cellular viability as well (Cedar & Bergman, 2011; Maunakea *et al.*, 2010; Jones, 2012). Its role however is even less clear than the one of gene promoter methylation.

Gene body methylation for example has been associated to increased expression levels, (Jones, 2012), and there is evidence that it is involved in the control of splicing (Lev Maor *et al.*, 2015; Jones, 2012). Thus, the relationship between DNA methylation and transcription is strongly dependent on the particular genomic and cellular context, and much more complex than was thought at first sight.

Furthermore, the epigenome interacts with the genome and vice versa (Zaina *et al.*, 2010; Oakes *et al.*, 2014). DNA damage for example can also cause epigenetic changes (Kovalchuk & Baulch, 2008), and methylation alterations are known to cooperate with mutational events in carcinogenesis (Jones, 2012; Chan *et al.*, 2008; Baylin & Jones, 2011; Oakes *et al.*, 2014). The long known global hypomethylation of tumors, first reported by Feinberg & Vogelstein (1983), has been associated to enhanced genomic instability (Jones & Baylin, 2002; Witte *et al.*, 2014), and it has been demonstrated for several cancers that mutations in methyltransferases lead to loss of DNA methylation which is followed by chromosomal instability (Qu *et al.*, 1999; Rodriguez *et al.*, 2006; Eden *et al.*, 2003).

Finally, the identification of a wide number of genes with aberrant methylation patterns in cancer led to the establishment of an enormous amount of methylation biomarkers for diagnosis, risk prognosis, and prediction of therapy response (Baylin & Jones, 2011; Witte *et al.*, 2014).

## 1.3 Epigenetic and Transcriptional Variability

### 1.3.1 Introduction to Biological Variability

"Life is a study in contrasts between randomness and determinism", stated Raj & Van Oudenaarden (2008). As previously described, genetically identical cells, or organisms, are able to obtain an incredible variety of phenotypes, even in completely homogenous environments. Gärtner (1990) showed that, although trying for more than 20 years, it was not possible to reduce phenotypic variability in animal inbreeding by controlling laboratory settings, and referred to this phenomenon as the "third component". That is, neither genetic variability nor the environment could explain the observed phenotypic diversity. Since then, many studies suggested that this additional source of diversity arises from randomness and noise in biological processes such as gene expression (Elowitz *et al.*, 2002; Raj *et al.*, 2010; Dong *et al.*, 2011; Kaern *et al.*, 2005).

Indeed, biological noise has emerged as an important factor influencing phenotypic variability. The first experiments measuring variability in the expression of a gene in *Escherichia coli* via the introduction of two copies of the same promoter into the genome highlighted the presence of two different causes of fluctuations in gene expression: intrinsic noise and extrinsic noise (Elowitz *et al.*, 2002; Swain *et al.*, 2002). The difference between the two types of noise is illustrated in figure 1.3.

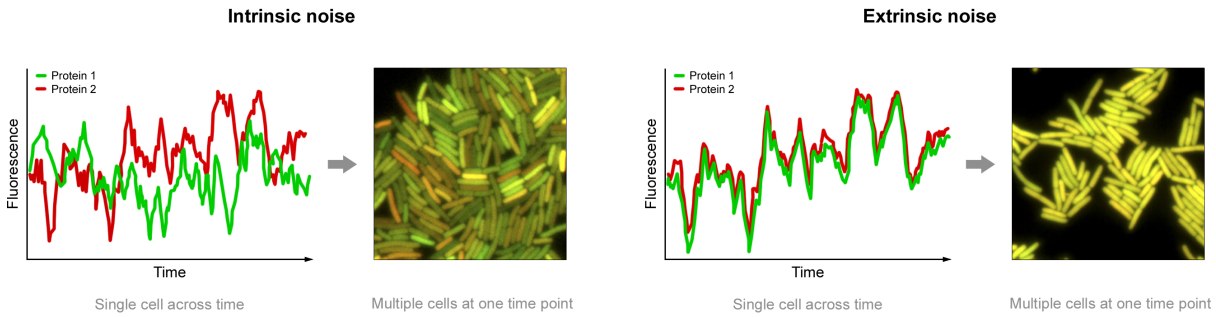


Figure 1.3: Intrinsic and extrinsic noise. Elowitz *et al.* (2002) constructed *Escherichia coli* strains by integrating two reporter genes (shown in green and red respectively) controlled by identical promoters. Cells with the same amount of protein appear yellow, cells expressing more of one of the two fluorescent proteins appear in red or green color shades. The expression of the two proteins can become uncorrelated in individual cells because of intrinsic noise, giving rise to a population in which some cells express more of one fluorescent protein than the other. When only extrinsic but no intrinsic noise is present, the two fluorescent proteins fluctuate in a correlated fashion over time, thus, each cell will have the same amount of both proteins at a given time point.
Figure adapted from Elowitz *et al.* (2002).

Extrinsic noise are fluctuations that originate from variabilities in external factors such as the environment (e.g. temperature or pressure), but can also relate to cell-cycle stage, cell

size, or mitochondrial content, for example. Thus, they affect the expression of all genes equally in a single cell, but may also be different from cell to cell, or over time (Elowitz *et al.*, 2002; Raser & O'Shea, 2005; Swain *et al.*, 2002). Extrinsic noise is a fundamental source of heterogeneity in prokaryotes and eukaryotes (Guantes *et al.*, 2015), and is probably composed of both stochastic and deterministic influences on the cell (Snijder & Pelkmans, 2011).

Intrinsic fluctuations in contrast are those that arise due to randomness inherent in biochemical processes in the cell, such as transcription and translation. They are affecting each copy of a gene independently. Intrinsic stochastic effects become especially prominent when there are only a few molecules of a specific type present in a cell (Elowitz *et al.*, 2002; Swain *et al.*, 2002).

An important source of intrinsic fluctuations is transcriptional and translational "bursting". It has been observed that variability in the expression of a gene depends on the rates of its transcription and translation (Ozbudak *et al.*, 2002). Proteins and also mRNA molecules are often produced at high frequency in short bursts, which are followed by quiescent periods, switching the gene randomly on and off for transcription (Raj & Van Oudenaarden, 2008; Lubeck & Cai, 2012; Raj *et al.*, 2006; Suter *et al.*, 2011). These bursts have especially been investigated in yeast and mammalian cells, and have been related to chromatin organization, where transcriptionally silenced heterochromatin leads to random events of gene activation and inactivation (Raj & Van Oudenaarden, 2008).

### 1.3.2 Biological Significance and Functions of Variability

Gene expression variability has of course important consequences for cellular function. Genes that are essential for the functioning of a cell like housekeeping genes which are responsible for protein synthesis, cell growth and general metabolism, to name a few examples, require stable and precise expression levels, and indeed, housekeeping genes have been shown to exhibit less variability than other classes of genes (Raj & Van Oudenaarden, 2008; Alemu *et al.*, 2014; Basehoar *et al.*, 2004; Li *et al.*, 2010).

In contrast, genes involved for example in stress-response tend to be highly variable, enabling a rapid adaptation of organisms to changing environmental conditions (Blake *et al.*, 2006; Dong *et al.*, 2011; Kaern *et al.*, 2005; Alemu *et al.*, 2014; Hulse & Cai, 2013), and thus leading to benefits in survival. This is, because it is easier to achieve large changes in gene expression in response to signaling if a gene already displays large stochastic fluctuations in absence of the stimulus, an observation is reminiscent of the "fluctuation

dissipation theorem". The theorem states that the response of a variable to perturbation is proportional to the fluctuation of that variable in absence of an applied force, thus, the more something varies under normal conditions, the more it will respond to perturbation (Lehner & Kaneko, 2011). See also figure 1.4 for a demonstration of the concept.

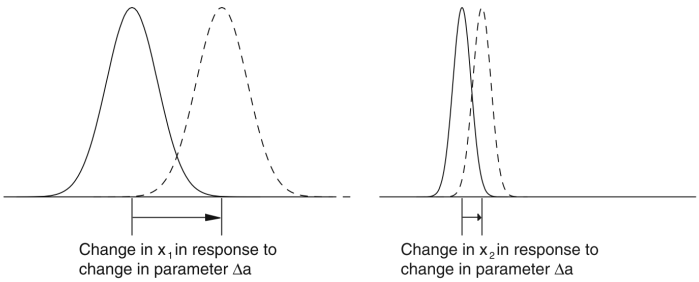


Figure 1.4: Schematic representation of the fluctuation dissipation theorem. The distribution of a phenotypic trait $x_1$ with large variance (left) shifts more than that of phenotype $x_2$ presenting a smaller variance (right) in response to a change in parameter $a$. Thus, the response of each trait is proportional to its fluctuation.

Figure adapted from Lehner & Kaneko (2011).

It is not only that genes with higher expression variability are more responsive to external stimuli, but there also exists a correlation between expression variation and the responsiveness to mutation, leading to a faster speed of evolution (Lehner & Kaneko, 2011). Expression variability is subject to evolutionary pressures and therefore linked with evolvability of complex organisms (Lehner, 2008; Kaern *et al.*, 2005; Hulse & Cai, 2013). So for any particular gene, there exists a relationship between its level of variability, its responsiveness to perturbation, and the potential to evolve (Lehner & Kaneko, 2011).

This finally leads to a strong correspondence between expression variability due to stochastic processes in single cells from the same population, and expression variability of single cells across different conditions or time points (Lehner & Kaneko, 2011). The equivalence between measuring variability at one time point in a population of for example 1,000 genetically identical cells, and measuring the variability of one single cell at 1,000 time points, is also known as the "ergodic hypothesis", often practically used to gain information about the nature of fluctuations at the single cell level by measuring an ensemble of cells (Brock *et al.*, 2009). Taking these considerations one step further, it has been shown that heterogeneity observed within populations correlates with heterogeneity measured across populations, and – although to a lesser extent – even across species (Lehner & Kaneko, 2011; Dong *et al.*, 2011; Tirosh *et al.*, 2009; Choi & Kim, 2009; Li *et al.*, 2010).

Besides the mentioned stress-response and signaling, stochastic gene expression is known to play a key role in development and cellular differentiation in multicellular organisms (Raj & Van Oudenaarden, 2008; Alemu *et al.*, 2014), by allowing for selection and propa-

gation of cell type-specific gene expression (Kaern *et al.*, 2005). This has been particularly investigated in the context of hematopoiesis and the immune system (Enver *et al.*, 1998; Hume, 2000; Chang *et al.*, 2008).

Immune cells exhibit extensive genetic variation, especially T and B cells. To be able to respond to a broad range of different antigens, they utilize DNA sequence recombination to generate diverse cell surface receptors (Satija & Shalek, 2014; Feinerman *et al.*, 2008). Beside this high variability in their receptor sequences, they also show genetic and non-genetic variability in further signaling molecules and key transcription factors, all necessary to generate diverse and effective immune responses (Paszek *et al.*, 2010; Busslinger & Tarakhovsky, 2014).

Epigenetic diversity plays an important role here (Pujadas & Feinberg, 2012). Epigenetic modifications have not only been linked to gene expression changes as described previously in section 1.2, but may also present stochastic fluctuations themselves (Kaern *et al.*, 2005). This diversity can facilitate fitness-enhancing alterations in changing conditions (Landau *et al.*, 2014b), as epigenetic states are known to be readily susceptible to environmental changes (Richards, 2006), and contributes to the plastic gene expression and evolutionary landscape in development and differentiation (Pujadas & Feinberg, 2012).

An important additional aspect that has to be taken into account when dealing with gene expression variability is that variation in gene expression levels does not necessarily affect the phenotype. This is often referred to as "robustness" (Paszek *et al.*, 2010; Barkai & Leibler, 1997; Lehner & Kaneko, 2011; Kellogg & Tay, 2015). Furthermore, robustness at the more global level of biological systems is often achieved through cell to cell coordination. For example, considering the tight control of programmed cell death in normal cells it seems unlikely that variation in timing and probability of apoptosis are a consequence of unstable regulation. Instead, by turning the strict binary decision (apoptosis – yes or no) at the single cell level into a graded response of the population of cells, variability is again likely to provide an adaptive advantage (Paszek *et al.*, 2010).

However, gene expression noise can also be undesirable. For example it has been shown that aging is correlated with an increased level of variability in both gene expression and also DNA methylation patterns (Raj & Van Oudenaarden, 2008; Li *et al.*, 2010; Fraga *et al.*, 2005; Southworth *et al.*, 2009; Somel *et al.*, 2006; Bahar *et al.*, 2006; Hannum *et al.*, 2013). Furthermore, hypervariable gene expression has been linked with human disease (Alemu *et al.*, 2014; Ho *et al.*, 2008; Prieto *et al.*, 2006; Feinberg *et al.*, 2010).

While the numbers of both up- and downregulated genes are often similar in classical differential expression analyses comparing mean levels of expression between disease and control samples – that is, the number of significant results is similar in both directions of regulation – it has been shown that gene expression variability instead is predominantly increased in diseased patients (Ho *et al.*, 2008).

Gene expression variability plays an important role in human immunodeficiency virus (HIV) susceptibility which is known to greatly vary across individuals (Li *et al.*, 2010), in neurological disorders (Li *et al.*, 2010; Mar *et al.*, 2011), and it has been strongly associated with cancer, where it has moreover been recently shown to provide useful and previously unseen information for diagnostic and predictive purposes (Bravo *et al.*, 2012; Marusyk *et al.*, 2012).

### 1.3.3 Variability in Cancer

Traditionally, tumor heterogeneity at different levels – such as intratumoral cell to cell variability, or heterogeneity across samples or individuals – has been explained by genetic variability due to random mutations and clonal evolution (Brock *et al.*, 2009; Lengauer *et al.*, 1998; Gerlinger *et al.*, 2012; Anderson *et al.*, 2011). This view has been challenged with the development of the cancer stem cell hypothesis, stating that tumorigenesis is driven by stem cell like cancer cells with indefinite self-renewal potential (Reya *et al.*, 2001). According to this theory, tumor heterogeneity arises from variable differentiation states of these cells (Marusyk *et al.*, 2012), and not from random genetic mutations (Brock *et al.*, 2009).

Given that epigenetic modifications are heritable to daughter cells, and therefore subject to natural selection, the contribution of epigenetic modifications to cancer is probably substantial, especially because epigenetic alterations accumulate as the cell population evolves, and diversifies at rates that are orders of magnitude higher than those of somatic genetic alterations (Landau *et al.*, 2014a). As increased epigenetic heterogeneity results in a more plastic evolutionary landscape, it facilitates the emergence of both genetic and epigenetic alterations that enhance fitness (Landau *et al.*, 2014b). Now, it is well known that epigenetic variability is contributing significantly to tumor heterogeneity (Hansen *et al.*, 2011; Brock *et al.*, 2009; Landau *et al.*, 2014b; Feinberg & Irizarry, 2010; Issa, 2011).

In all cancers ever investigated in terms of epigenetic variability so far, a strong increase of variation in tumor samples compared to healthy tissue-matched normal ones has been observed (Timp & Feinberg, 2013; Pujadas & Feinberg, 2012; Hansen *et al.*, 2011; Jaffe

*et al.*, 2011), similar to the observations made when looking at gene expression variability (see above). Additionally, the difference in DNA methylation variability between cancer and normal samples is strikingly higher than the mean differences in DNA methylation measured traditionally (Hansen *et al.*, 2011; Pujadas & Feinberg, 2012).

Also in cancer, genes with hypervariable DNA methylation patterns, as well as those exhibiting increased variability in gene expression, are associated with cellular differentiation, development, mitosis and cell cycle (Pujadas & Feinberg, 2012; Hansen *et al.*, 2011; Bravo *et al.*, 2012). Furthermore, sites with increased DNA methylation heterogeneity have been associated to genes that are tissue-specific in normal samples (Pujadas & Feinberg, 2012; Hansen *et al.*, 2011), but interestingly not specifically expressed in the normal tissue of the corresponding cancer (Bravo *et al.*, 2012; Alemu *et al.*, 2014), indicating a deregulation of particular tissue-specific genes in cancer.

Importantly, stochastic heterogeneity is also linked to therapeutic resistance (Marusyk *et al.*, 2012). The probably most commonly known example is the one of bacterial resistance after treatment with antibiotics (Balaban *et al.*, 2004). Although most of the bacterial population is killed by antibiotic treatment, a small subset of so called "persistor" cells can survive, enabling the reemergence of the infection after the treatment has been stopped (Kaern *et al.*, 2005). Stochastic mechanisms are thought to play a significant role in these phenomenons (Raj & Van Oudenaarden, 2008).

Similar effects have been observed in cancer cells treated with chemotherapeutic agents (Brock *et al.*, 2009; Paszek *et al.*, 2010). In fact, many cancer drugs show so called "fractional killing" (Berenbaum, 1972), in which each round of therapy kills some but not all cancerous cells (Spencer *et al.*, 2009). Cohen *et al.* (2008) showed that cell to cell variability increases after drug addition, with dramatic differences in the dynamics of the expression of cell death related proteins, allowing some cells to escape from the treatment. Such effects have also been observed in human cell lines after tumor necrosis factor related apoptosis-inducing ligand (TRAIL) exposure, where the ability of some cells to survive apoptosis was linked to noise-driven differences in the levels of proteins that regulate receptor-mediated cell death.

As such heterogeneity within cancer cell populations leads to relapse with the outgrowth of resistant cancer cells after initial treatment (Cohen *et al.*, 2008; Gascoigne & Taylor, 2008), and the underlying mechanisms seem often to be non-genetic (Marusyk *et al.*, 2012), there is a great potential for drugs reducing this diversity via epigenetic modifications modulating cellular heterogeneity (Sharma *et al.*, 2010; Marusyk *et al.*, 2012; Paszek

*et al.*, 2010). It is clear that cell to cell variability and changes in cellular phenotypes resulting from adaptation to treatment and selection for resistant phenotypes need to be taken into account in order to overcome resistance and improve therapeutic outcome in cancer treatment (Marusyk *et al.*, 2012).

## 1.4 Human Blood Cells

Blood contains cells with diverse functions, from the transport of nutrients and oxygen to the building of antibodies and killing pathogens. All blood cells originate from a pluripotent hematopoietic stem cell, located mainly in the bone marrow, respectively in the liver in fetuses (Alberts *et al.*, 2004). These stem cells divide infrequently to build new stem cells (self-renewal) and determined precursor cells that further divide and differentiate into mature blood cells (Alberts *et al.*, 2004).

The three main components of the blood are red blood cells (erythrocytes), platelets, and white blood cells (leukocytes), where the leukocytes are further subdivided into myeloid and lymphoid cells, see figure 1.5.

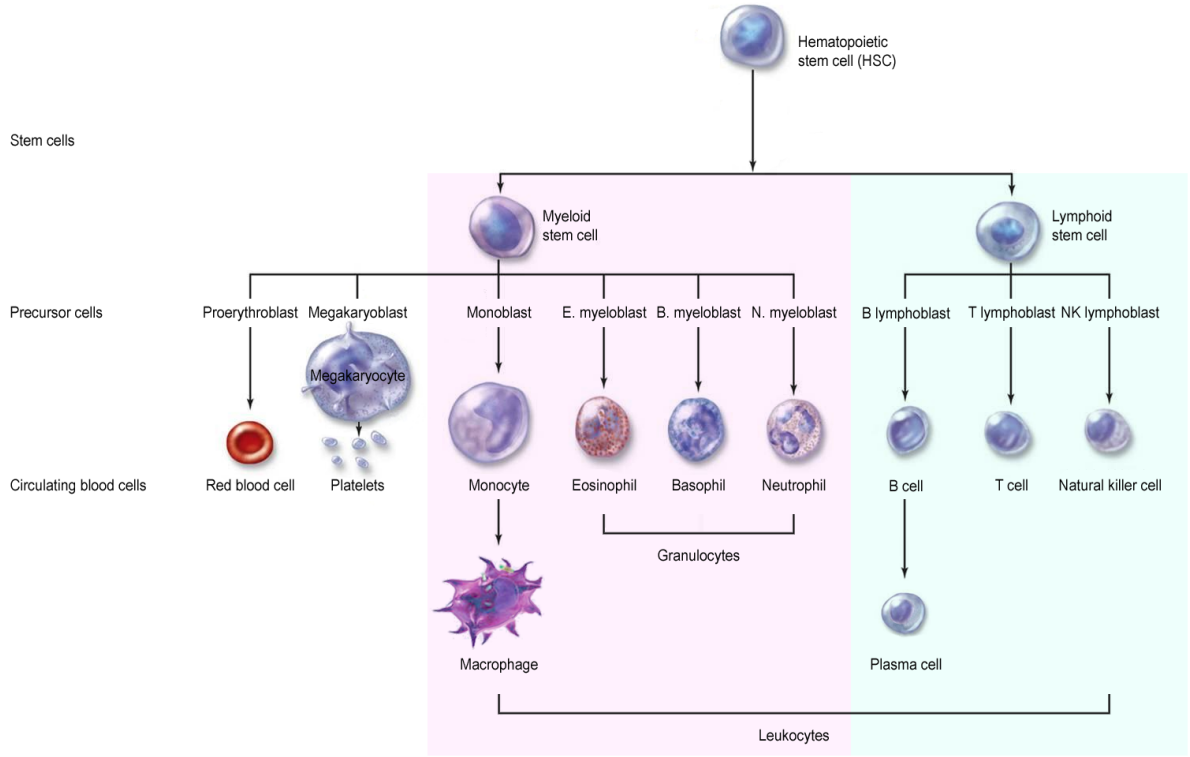


Figure 1.5: Hematopoiesis. The pluripotent hematopoietic stem cell can replicate and differentiate into a myeloid or lymphoid stem cell, which can further differentiate into intermediate progenitor cells that then differentiate into mature blood cells. The myeloid and lymphoid compartment are background colored in light magenta and cyan.
Figure adapted from OpenStax College (2013).

Leukocytes are cells of the immune system – described in more detail in my Bachelor thesis (Ecker, 2009) – responsible for fighting infection and the phagocytosis of foreign invaders and cell parts.

The most abundant cell type among leukocytes are neutrophils, which belong to the group of granulocytes. They account for 30 to 80% of all leukocytes (Stemcell Technologies, 2015) and live only a few hours to days (Alberts *et al.*, 2004; Summers *et al.*, 2010; Wheater *et al.*, 1979). Neutrophils have a multilobular nucleus and are therefore also called polymorphonuclear cells. They ingest microorganisms, in particular bacteria, and are essential in the innate immune system to fight infection. Neutrophils are among the first responders that migrate to sites of inflammation, and they can release proteins (granules) to combat infection (Borregaard, 2010). Moreover, they can form so-called neutrophil extracellular traps (NETs), which are webs of fibers composed of chromatin and granule proteins that trap and kill pathogens extracellularly (Brinkmann *et al.*, 2004).

The largest leukocytes are monocytes. They live longer than neutrophils, approximately one to five days (Alberts *et al.*, 2004; Geissmann *et al.*, 2010; Wheater *et al.*, 1979; Pillay *et al.*, 2010; Kolaczkowska & Kubes, 2013). Monocytes migrate from the blood stream to other tissues where they mature into macrophages, dendritic cells, and other cell types (Gordon & Taylor, 2005). Together with neutrophils, they are the phagocytosis experts of the innate immune system (Hoffbrand *et al.*, 2005), and they constitute 2 to 12% of all leukocytes (Stemcell Technologies, 2015).

Within lymphocytes, which are mostly found in the lymphatic system where they proliferate, there exist two further important groups of cells, both crucial for the adaptive immune system: B cells, which produce antibodies that can bind to pathogens ("antibody-mediated immunity") among other functions, and T cells, which help coordinating the immune response ("cell-mediated immunity") and kill virus infected cells (Alberts *et al.*, 2004). The third group of lymphocytes are natural killer (NK) cells, and as their name says, they can kill infected and also some cancerous cells.

One of the most abundant T cell types is the naive $CD4^+$ cell, comprising around 1 to 10% of all leukocytes (Stemcell Technologies, 2015). Naive $CD4^+$ T cells are mature cells that have not yet encountered their antigen. They are long-lived (living weeks to years), and their life span increases with age (Tsukamoto *et al.*, 2009). $CD4^+$ T cells are capable of responding to novel pathogens, get activated when their antigen binds to the T cell receptor (TCR) and they help antigen-presenting cells through cell to cell interactions and the secretion of cytokines (Alberts *et al.*, 2004).

Unstimulated B cells and T cells are morphologically very similar, they are both small and almost completely filled with their nuclei (Alberts *et al.*, 2004). B cells can be distinguished from T cells and NK cells by the presence of a different receptor, the B cell receptor (BCR), allowing a B cell to bind to a specific antigen. B cell development, as T cell development, occurs through several stages. For T cells this process is described in my Bachelor thesis (Ecker, 2009).

Each stage of B cell development involves changes in the genome content of the antibody loci (Janeway *et al.*, 2001). Five classes of antibodies (immunoglobulins) exist in mammals: IgA, IgD, IgE, IgG, and IgM (Alberts *et al.*, 2004). Antibodies are composed of two identical light (L) and heavy (H) chains. The genes specifying these chains are found in the variable (V) and constant (C) region (Alberts *et al.*, 2004). In early B cell development, every B cell forms a distinct BCR through random combinations in these regions (Alberts *et al.*, 2004; Zenz *et al.*, 2010). In the heavy-chain V region there are three segments of recombinations, namely V, D, and J. In the light chain only two segments are involved in the rearrangement, V and J (Alberts *et al.*, 2004). The random combinations allow B cells to generate a tremendous diversity of potential BCRs (Janeway *et al.*, 2001; Alberts *et al.*, 2004; Blachly *et al.*, 2015).

The BCR repertoire is further increased by the process of somatic hypermutation activated by antigen-binding (Zenz *et al.*, 2010; Alberts *et al.*, 2004). Activated B cells enter B cell follicles in secondary lymphoid organs, called germinal centers (GCs), where they undergo massive clonal expansion, accompanied by the process of somatic hypermutation which modifies the immunoglobulin variable region genes by introducing mutations into them at a high rate (Klein & Dalla-Favera, 2008; Zenz *et al.*, 2010). Affinity-increasing mutations are selected, and positively selected B cells normally undergo multiple rounds of proliferation and mutational selection until they finally differentiate into memory B cells or plasma B cells and leave the GC (Zenz *et al.*, 2010; Alberts *et al.*, 2004). Many GC B cells also undergo class-switch recombinations. B cells with unfavorable mutations die through apoptosis in the GC (Zenz *et al.*, 2010).

## 1.5 Chronic Lymphocytic Leukemia

### 1.5.1 The Disease of Chronic Lymphocytic Leukemia

Chronic lymphocytic leukemia (CLL) is the most frequent leukemia in adults in Western countries (Rozman & Montserrat, 1995; Caligaris-Cappio & Hamblin, 1999; Zenz *et al.*, 2010). The disease normally affects elderly people, with a median age of around 72 years

at diagnosis (Zenz *et al.*, 2010). It rarely occurs in children, and the incidence is higher in men than in women (Rozman & Montserrat, 1995; Zenz *et al.*, 2010).

In CLL, abnormal B lymphocytes (shown in figure 1.6) which are not able to fight infection well accumulate in the blood due to the inhibition of cell death (Rozman & Montserrat, 1995; Caligaris-Cappio & Hamblin, 1999). As the number of leukemic cells increases in the blood and bone marrow, there is less room for healthy leukocytes, red blood cells and platelets. Over time, the neoplastic cells can spread to other parts of the body, including lymph nodes, liver, and spleen (Rozman & Montserrat, 1995).

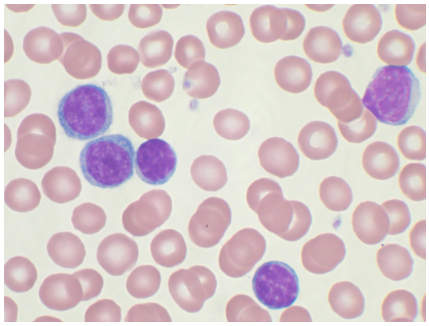

Figure 1.6: CLL cells. High-power magnification (1000 X) of a stained peripheral blood smear. The lymphocytes with the darkly staining nuclei and scant cytoplasm are CLL cells.
Image taken from Thompson (2006).

The reasons for the development of CLL have not yet been fully elucidated. The disease is thought to arise due to several pathogenic mechanisms involving microenvironmental stimuli as well as genetic and epigenetic events (Guarini *et al.*, 2008; Zenz *et al.*, 2010).

Of particular importance in CLL is the BCR, as it is believed that disease onset and progression is driven by the antigenic stimulation of the BCR and/or cell-autonomous BCR signaling (Chiorazzi & Ferrarini, 2003; Quiroga *et al.*, 2009; Iacovelli *et al.*, 2015). CLL cells exhibit stereotyped BCRs across patients (Dühren-von Minden *et al.*, 2012), supporting the hypothesis that the recognition of specific antigens drives CLL pathogenesis and evolution of the disease (Zenz *et al.*, 2010; Herishanu *et al.*, 2011; Dühren-von Minden *et al.*, 2012). Consequently, as signaling by antigens and the BCR can influence the clinical course of CLL, it has been suggested as a therapeutic target (Herishanu *et al.*, 2011), and the inhibition of BCR signaling seems to be a promising approach in early clinical trials indeed (Friedberg *et al.*, 2010; Shain & Tao, 2013).

However, the diagnosis of CLL does not directly imply the need for therapy, as it does usually not cause symptoms at early stages and is often only found during routine blood tests returning abnormally high white blood cell counts (Rozman & Montserrat, 1995; Zenz *et al.*, 2010). Such patients in early stages of the disease get monitored closely, but

only problems caused by the disease, such as infection, are treated, and not the leukemia itself. Treatment of CLL gets started when the disease has progressed to a point where it may affect the patient's quality of life, when constitutional symptoms such as bulky lymphadenopathy or splenomegaly appear, and when specific markers according to clinical staging systems are met (Rozman & Montserrat, 1995). The current gold standard in patient evaluation and treatment decision in CLL are the staging systems of Rai *et al.* (Rai *et al.*, 1975; Rai & Han, 1990) and Binet *et al.* (1981).

Although CLL is considered incurable, later stages of the disease can be treated by different options including radiation therapy, chemotherapy, surgery (removal of the spleen), stem cell transplantation, and monoclonal antibody therapy (also called biotherapy) given by infusion (Rozman & Montserrat, 1995).

However, CLL generally progresses slowly in most cases. Due to the often indolent course, older patients with early and stable disease may not need any treatment in their lifetimes and survive as long as normal subjects of the same age (Rozman & Montserrat, 1995). Early therapeutic intervention does not improve survival time or quality of life of the patients (French Cooperative Group on Chronic Lymphocytic Leukemia, 1990; Montserrat *et al.*, 1991; Catovsky *et al.*, 1991; CLL Trialists' Collaborative Group, 1999) and has even been associated with shorter survival (Montserrat *et al.*, 1991).

The prognosis of CLL is however highly variable, and there exist subtypes of CLL with very different clinical outcomes. Important prognostic parameters are for example the expression of the intracellular protein *ZAP70* (zeta-chain-associated protein kinase 70) and the membrane glycoprotein *CD38* (cluster of differentiation 38). CLL that is positive for these markers shows decreased average survival (Zent & Kay, 2007; Rassenti *et al.*, 2008).

Additionally, CLL prognosis is dependent on genetic changes within the neoplastic cell population. The major genetic aberrations impacting clinical outcome in CLL are the following (Döhner *et al.*, 2000; Rozman & Montserrat, 1995):

- Deletions of chromosome 17 which target the cell cycle regulation protein *P53* (tumor protein 53). Patients with this abnormality have a significantly shorter time to treatment and show often poor response to conventional drug therapy (Fabris *et al.*, 2008). The deletion is found in 5–10% of the patients.
- Deletions on chromosome 11 that target the *ATM* (ataxia telangiectasia mutated) gene. The deletion is unfavorable and affects 5–10% of the patients.
- An additional chromosome 12. It is found in 20–25% of the patients and leads to an intermediate prognosis (Juliusson & Gahrton, 1993; Gahrton *et al.*, 1980).

- Deletion at band q14 of chromosome 13 is the most common one in CLL. The region targets the *RB1* (retinoblastoma 1) gene (Rozman & Montserrat, 1995; Fält *et al.*, 2005) and a microRNA cluster (Mraz *et al.*, 2009) that functions as a tumor suppressor, with oncogene *BCL2* (B cell CLL/lymphoma 2) as its target (Bonci *et al.*, 2008). Patients with this deletion have a more favorable prognosis. About 50% of patients have this defect in their CLL cells.
- Deletions of chromosome 6 and 11 (Juliusson & Gahrton, 1993).

The most reliable and best-studied diagnostic parameter in CLL is the mutation status of the immunoglobulin sequence (Zent & Kay, 2007), leading to two subgroups of patients with different clinical courses (Hamblin *et al.*, 1999; Zenz *et al.*, 2010; Damle *et al.*, 1999; Chiorazzi & Ferrarini, 2003).

The immunoglobulin variable-region heavy chain ($IgV_H$) gene mutation status reflects the maturity of the lymphocytes, as the $IgV_H$ somatic hypermutation is a physiologic marker of antigen exposure and passage through the germinal center (Zent & Kay, 2007; Rosenwald *et al.*, 2001), see also section 1.4. Increased somatic mutation rates in the corresponding region – that is 2% or greater difference from the germline sequence – indicate mature lymphocytes, and CLL patients showing $IgV_H$ gene mutations have a significantly better prognosis (Hamblin *et al.*, 1999), presenting a median survival of more than 24 years (Chiorazzi & Ferrarini, 2003). This subtype of CLL is referred to as $IgV_H$ "mutated" CLL (M-CLL). In contrast, $IgV_H$ "umutated" CLL (U-CLL) shows a more immature cell pattern with few mutations in the $IgV_H$ antibody gene region (Hamblin *et al.*, 1999) and U-CLL patients are high risk patients with worse prognosis and a median survival of 4–8 years (Chiorazzi & Ferrarini, 2003).

U-CLL patients are at average slightly older at diagnosis than patients with M-CLL (Hamblin *et al.*, 1999). The two groups M-CLL and U-CLL also show several further biological differences with implications for clinical outcome: different levels of *ZAP70* and *CD38* expression, differential activity of key signal transduction pathways, different telomere lengths, different proliferation capacity, and different likelihood of genetic lesions and mutations (Kröber *et al.*, 2002; Klein *et al.*, 2001; Rosenwald *et al.*, 2001; Stilgenbauer *et al.*, 2007; Puente *et al.*, 2011). In summary, U-CLL cells are more likely to show alterations associated to poor prognosis, whereas M-CLL shows higher proportions of changes with favorable clinical outcome (Chiorazzi & Ferrarini, 2003).

Further information about CLL, especially gene expression and DNA methylation in CLL and its subtypes, can be found in Annex 2.

### 1.5.2 Variability in Chronic Lymphocytic Leukemia

Tumor heterogeneity has been traditionally investigated in solid malignancies, with an emphasis on analyzing genetic variation and clonal evolution, as described in section 1.3.3. However, even CLL, previously thought to progress via monoclonal expansion (Rozman & Montserrat, 1995; Klein *et al.*, 2001) can display genetic variability, both at the level of single tumoral cells as well as through clonal heterogeneity (Stilgenbauer *et al.*, 2007; Landau *et al.*, 2013; Gurrieri *et al.*, 2002; Quesada *et al.*, 2011; Wang *et al.*, 2011; Schuh *et al.*, 2012; Landau *et al.*, 2014a; Gunnarsson *et al.*, 2011).

Increased genomic complexity has been correlated with decreased survival in CLL (Roos *et al.*, 2008; Ramsay *et al.*, 2013), and interestingly, clonal evolution seems to occur mainly in U-CLL cases (Stilgenbauer *et al.*, 2007; Herishanu *et al.*, 2011; Landau *et al.*, 2014a; Gunnarsson *et al.*, 2011) while M-CLL shows an increased number of (clonal, but not subclonal) somatic mutations compared to U-CLL (Landau *et al.*, 2013; Puente *et al.*, 2011; Quesada *et al.*, 2011).

As stated in section 1.3.3, genetic variability might have an impact on epigenetic and transcriptional variability and vice versa, and also in CLL epigenetic modifications are probably involved in the phenotypic differences observed. Indeed, when Landau *et al.* (2014b) analyzed DNA methylation data of CLL, they found higher intra-sample heterogeneity in the leukemia cases compared to normal B cell samples, resulting from an increased proportion of cells with variable methylation patterns in terms of discordant methylation states in neighboring CpGs, called "locally disordered methylation".

This disordered methylation appeared to arise from stochastic processes, and was significantly associated with a reduced correlation between promoter methylation and gene silencing. Additionally, Landau *et al.* (2014b) found higher levels of disordered methylation in promoters of genes that showed already increased methylation variability in normal B cells, and in promoters of genes that were transcriptionally silenced in both normal B cells and CLL.

The genes affected by locally discordant methylation were enriched for genes important to pluripotency potential such as stem cell modules. Samples with increased DNA methylation heterogeneity also exhibited higher numbers of subclonal mutations. Finally, increased heterogeneity at the level of DNA methylation was associated with shorter survival in the study of Landau *et al.* (2014b).

Oakes *et al.* (2014) reported in another publication that DNA methylation heterogeneity correlated with advanced genetic subclonal complexity, and that it occurred at higher levels in U-CLL cases, again suggesting that DNA methylation variability might be associated with a more aggressive disease.

Also in CLL, tumor heterogeneity is a crucial factor to consider in therapy, as treatment often seems to accelerate evolution from heterogeneity present before the start of the therapy to increased fitness and a more aggressive phenotype of the disease (Landau *et al.*, 2013). This is thought to occur by favoring the rapid growth of more aggressive clones which benefit from the removal of incumbent clones due to the treatment, supporting the view that not treating CLL in early indolent stages (see section 1.5) leads to better clinical results (Landau *et al.*, 2014a).

Finally, it has been concluded that diversity at any level, be it genetic or non-genetic, is sufficient to influence clinical outcome in CLL, and monitoring heterogeneity during disease course might be a beneficial strategy to improve therapeutic decisions and risk prediction in patients (Kleppe & Levine, 2014; Swanton & Beck, 2014; Oakes *et al.*, 2014; Landau *et al.*, 2013).

Variability at the level of gene expression has however not been investigated in CLL so far.

# Chapter 2

# Objectives

The aim of this work is to characterize interindividual epigenetic and transcriptional variability and their interrelationship in the human hematopoeitic system in both health and disease, with a special emphasis on the following two areas that are investigated in detail in this context:

- Chronic lymphocytic leukemia
- Normal blood cells

Within these two main areas of the present thesis, the specific aims for each of them are the following:

**1. Analysis of Variability in Chronic Lymphocytic Leukemia**

- Measuring differential gene expression variability between the two main CLL subtypes using robust methods to quantify variability taking the mean-variance relationship into account.
- Validating the findings on differential variability in independent publicly available CLL datasets.
- Analyzing the relation between gene expression variability and DNA methylation in CLL.
- Characterizing the functions of genes with differential expression variability.
- Predicting the disease subtype of patients based on expression variability measurements.

**2. Analysis of Variability in Normal Blood Cells**

- Establishing robust and comparable methods to quantify interindividual variability as well as mean differences in both DNA methylation and gene expression data able to deal with the complex relationship between mean and variability measurements.
- Analyzing differential DNA methylation variability across monocytes, neutrophils and T cells.
  - Identifying cell type specific sites with hypervariable DNA methylation patterns.
  - Identifying sites with hypervariable DNA methylation patterns shared between two of the three cell types.
  - Identifying sites with hypervariable DNA methylation patterns in common in all three cell types.
- Analyzing differential gene expression variability across monocytes, neutrophils and T cells.
  - Identifying cell type specific genes with hypervariable gene expression patterns.
  - Identifying genes with hypervariable gene expression patterns shared between two of the three cell types.
  - Identifying genes with hypervariable gene expression patterns in common in all three cell types.
- Analyzing sex-specific differential DNA methylation and gene expression within each cell type and its possible contribution to interindividual variability.
- Analyzing the relation between DNA methylation variability and gene expression.

Thus, the main focus of this writing lies on the analysis of differential variability, with a comprehensive analysis of differential gene expression variability between the two main subtypes of CLL, and the creation of robust methodology to analyze both differential DNA methylation variability and gene expression variability, applied on a dataset comprising three normal blood cell types – monocytes, neutrophils and T cells.

# Chapter 3

# Methods

## 3.1 Gene Expression Variability in CLL

The description of materials and methods given in this section is also published in a modified form in Ecker *et al.* (2015).

### 3.1.1 Material

The work on gene expression variability in CLL presented here was conducted within the framework of the ICGC (International Cancer Genome Consortium, 2010). The ICGC coordinates large-scale cancer genome studies of 50 different tumor types and subtypes of clinical and societal importance around the globe. The aim of the consortium is to obtain a comprehensive catalogue of genomic, transcriptomic and epigenomic alterations in these cancer types. Our group participates in the Spanish Consortium with its CLL Genome Project (International Cancer Genome Consortium, 2015), which aims to decipher the diversity and complexity of genomic, epigenomic and transcriptomic changes in the genome of CLL and its subtypes M-CLL and U-CLL with the ultimate goal to improve prevention, diagnosis and treatment of the disease.

The first publications of the ICGC CLL Genome Project described recurrent mutations in CLL (Quesada *et al.*, 2011; Puente *et al.*, 2011). Subsequent studies focussed on the DNA methylome (Kulis *et al.*, 2012) and transcriptome (Ferreira *et al.*, 2014) of the disease and its subtypes, and gave rise to the here presented work. The preceding epigenomic and transcriptomic studies of CLL in which we also participated are summarized in Annex II where the corresponding papers we have published can be found as well.

For the study of gene expression variability in CLL presented in this thesis we used the ICGC CLL microarray datasets previously described in Kulis *et al.* (2012) and Ferreira *et al.* (2014), together with additional publicly available datasets for validation.

Gene expression measurements of the ICGC data were obtained by Affymetrix Human Genome U219 Array Plates by Kulis *et al.* (2012) and Ferreira *et al.* (2014). A total of 48,786 features of the microarray passed quality controls and filtering. Raw files were preprocessed and normalized by Kulis *et al.* (2012) and Ferreira *et al.* (2014) using the robust multi-array average (RMA) algorithm (Irizarry *et al.*, 2003) and the Affy package (Gautier *et al.*, 2004). The dataset comprises 122 CLL samples (70 M-CLL and 52 U-CLL) and 20 control samples of different healthy B cells (five naive B cells, three $IgM^+$ and $IgD^+$ memory B cells, four $IgA^+$ and $IgG^+$ memory B cells, and eight $CD19^+$ Bcells).

For the validation of our results, we included an additional gene expression dataset of CLL published by Fabris *et al.* (2008) under gene expression omnibus (GEO) accession number GSE9992, containing 60 samples (24 M-CLL and 36 U-CLL) and 22,215 probes in our analyses. The microarray platform used in this study was the Affymetrix Human Genome U133A Array. The data were quality assessed and preprocessed independently from the ICGC gene expression dataset. For normalization we used the fRMA algorithm (McCall & Irizarry, 2011).

To further confirm the main results of our analysis we used data published by Haslinger *et al.* (2004). This dataset is available under GEO accession number GSE2466 and we analyzed the 39 samples of M-CLL and the 33 U-CLL samples which were hybridized onto the Affymetrix Human Genome U95 Version 2 Array containing 12,625 probes. The dataset was normalized using the RMA algorithm (Irizarry *et al.*, 2003).

DNA methylation was measured by Infinium Human Methylation450K BeadChips. A total of 282,470 probes (139,076 of them falling into gene promoter regions) passed quality control and filtering procedures of Kulis *et al.* (2012). The data were analyzed by Genome Studio (Illumina, Inc.) and R using the lumi package (Du *et al.*, 2008), and an optimized analysis pipeline was developed and applied by Kulis *et al.* (2012). This pipeline includes several filters to exclude technical and biological biases that might produce false results, removing probes with low detection p-values, sex-specific and individual-specific methylation, or overlapping with single nucleotide polymorphisms (SNPs). To correct for the differing performance of Infinium I and Infinium II assays, subset-quantile within array normalization (SWAN) was applied (Makismovic *et al.*, 2012).

### 3.1.2 Measuring Gene Expression Variability

We used two different measures to quantify gene-wise expression variability. Firstly, we calculated the coefficient of variation (CV) of every gene $i$, defined as the ratio between

the sample standard deviation $s_i$ of expression values across patients and the sample mean $\bar{x}_i$, see formula 3.1.

$$CV_i = \frac{s_i}{\bar{x}_i} \tag{3.1}$$

Secondly, we quantified expression variability using the expression variability score introduced by Alemu *et al.* (2014), subsequently called EV. Alemu *et al.* (2014) applied local polynomial likelihood estimation Loader (1999) to model variance as a function of the mean of expression. In some more detail, the method assumes that variance is gamma distributed and estimates the expected variability of a gene given its overall expression values by a gamma regression model using a locally weighted quadratic polynomial. The ratio of observed variance to expected variance gives then the measurement of expression variability for each gene.

To compare gene expression variability between M-CLL and U-CLL and identify the top genes with differential variability, we calculated gene-wise CV differences $CVdiff_i = CV_{i,M\text{-}CLL} - CV_{i,U\text{-}CLL}$ and EV differences $EVdiff_i = EV_{i,M\text{-}CLL} - EV_{i,U\text{-}CLL}$.

To assess statistical significance, we performed gene-wise F-tests (Snedecor & Cochran, 1989) comparing M-CLL with U-CLL using R's `var.test()` function (R Development Core Team, 2008). Multiple hypotheses testing correction was performed using the Benjamini-Hochberg algorithm (Benjamini & Hochberg, 1995).

### 3.1.3 Analysis of DNA Methylation and its Relationship to Gene Expression Variability

To investigate the relationship between gene expression and DNA methylation, we mapped the microarray probe identifiers to Ensembl identifiers and used the average of the measurements for each gene. DNA methylation features were mapped to genomic regions using annotation information provided by Illumina, see Kulis *et al.* (2012). We applied the bumphunter method (Jaffe *et al.*, 2011) to identify regions of differential methylation between M-CLL and U-CLL. Smoothing of methylation values was applied and 1,000 permutations were performed to assess statistical significance.

Subsequently, we looked at the genomic annotation of the microarray probes within the regions which had been identified to be differentially methylated and assigned all regions to be either promoter regions or gene body regions if they contained at least three probes of the corresponding annotation. Regions not containing the described minimum of three probes were excluded from further analyses.

To detect if genes with their promoters or gene bodies lying within differentially methylated regions were significantly enriched in genes with increased variability in U-CLL we performed hypergeometric tests for both hyper- and hypomethylated regions using the R function `phyper()`. The test was performed on the basis of the 15,037 genes in common between the DNA methylation and gene expression data we used.

### 3.1.4 Functional Analysis

To test for enrichment of biological functions and pathways we used DAVID (Huang *et al.*, 2009). We uploaded the list of the top 500 genes of the ICGC CLL dataset and the top 500 genes of the Fabris CLL dataset and used the corresponding set of genes analyzed in the dataset as background set. We tested for the following functional annotation: GOTERM_BP_ALL, GOTERM_CC_ALL, GOTERM_MF_ALL, KEGG_PATHWAY, and REACTOME_PATHWAY and set the threshold of counts to a minimum of three genes. We considered terms and pathways as significantly enriched when the corresponding p-value adjusted by the Benjamini-Hochberg algorithm for multiple hypotheses correction (Benjamini & Hochberg, 1995) was smaller than 0.05.

The same analyses were performed in R using the packages GOstats (Falcon & Gentleman, 2007) and Category (Gentleman, 2015), and very similar results were obtained (data not shown).

The enrichment analyses on the five network modules were performed the same way as described above, except the background gene set used, which in this case was the set of all genes contained in the entire B cell network of Lefebvre *et al.* (2010) that are also present on the microarray platforms investigated ($n$ = 5,548).

### 3.1.5 Random Forest Classification

We applied the randomForest R package (Liaw & Wiener, 2002) to create random forest classifiers and the package ROCR (Sing *et al.*, 2005) to calculate area under the curve (AUC) values, which were used to evaluate the prediction of the disease subtype of the patients in our independent validation dataset.

### 3.1.6 Programming Language

If not stated otherwise, the analyses were performed using R version 3 (R Development Core Team, 2008) and Bioconductor (Bioconductor, 2015).

## 3.2 Variability in Normal Blood Cells

### 3.2.1 Material

We analyzed the DNA methylation microarray and RNA sequencing (RNAseq) pilot dataset generated by BLUEPRINT (Adams *et al.*, 2012) for its Human Variation Epigenome Project (see figure 3.1). The BLUEPRINT epigenome project is a large-scale European research effort in which our group participates. It is the European cornerstone of the International Human Epigenome Consortium (IHEC), an international research cooperation aiming to coordinate epigenome mapping for a broad spectrum of human cell types.

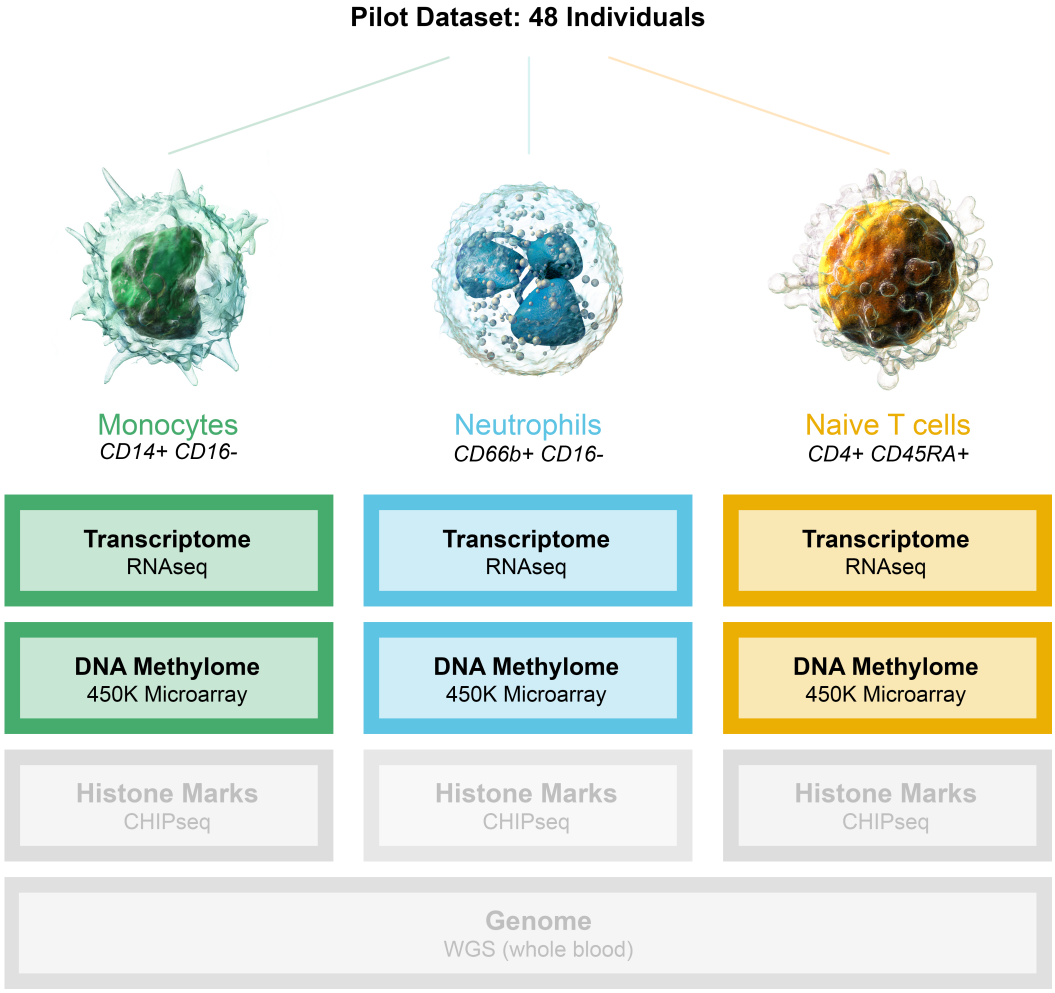


Figure 3.1: Overview of the dataset comprising the transcriptome and DNA methylome of monocytes, neutrophils and T cells derived from 48 individuals. In the future, the dataset will be extended to 200 individuals. Furthermore, histone marks will be added, and whole genome sequencing (WGS) data obtained from whole blood will also be available for all individuals. The colors of the three cell types represent the color scheme used in all subsequent figures comparing these cell types.
Images of cell types adapted from Blausen.com (2014).

The aim of the BLUEPRINT consortium is to generate at least 100 reference epigenomes of distinct types of human hematopoietic cells from healthy individuals and their malignant leukemic counterparts to advance the knowledge of biological processes and mechanisms in health and disease, systematically linking epigenetic variation with phenotypic plasticity in both health and disease.

The BLUEPRINT dataset used in this work comprises DNA methylation and gene expression data of three different blood cell types: monocytes, neutrophils and T cells (see section 1.4 for an introduction of these cell types). The samples were derived from 48 unrelated healthy individuals. An overview of the data can be found in figure 3.1. Altogether, the dataset contains samples derived from 34 male and 14 female individuals within an age range of 30 to 70 years. Data of T cells are only available for 40 of the 48 individuals.

In the future, the dataset will be extended to 200 individuals, and several histone marks will be added: H3K4me1 and H3K27ac for all three cell types, and additionally H3K27me3 for neutrophils. Also WGS data obtained from whole blood will be available for all individuals contained in the study.

### 3.2.2 Data Preprocessing and Filtering

The general preprocessing, quality assessment and filtering of the DNA methylation and gene expression data was done by others at University College London Cancer Institute and Wellcome Trust Sanger Institute respectively, and is therefore only summarized here for completeness.

DNA methylation measurements were obtained by Infinium Human Methylation450K BeadChips. Raw data (IDAT files) containing information for all 485,512 probes were normalized using minfi (Aryee *et al.*, 2014) and functional normalization (Fortin *et al.*, 2014), and batch-effect corrected by ComBat (Johnson *et al.*, 2007). Probes were filtered corresponding to the following criteria:

- probes with median detection p-value $> 0.01$ in more than one sample
- probes with bead count of less than three in more than 5% of samples
- probes mapping to sex chromosomes
- probes mapping to multiple locations with at least two mismatches
- non-CG probes
- probes with SNPs in European population based on 1000 Genomes Project phase I

A set of 423,089 probes passed the filtering procedure. Seven of the 40 DNA methylation samples of T cells were excluded from the analyses because of quality problems.

RNAseq data (100 bp single-end, total RNA) was quality controlled using FASTQC (Andrews, 2014). Sequencing reads were aligned to the GRCh37 human reference genome using STAR (Dobin *et al.*, 2013) and GSNAP (Wu & Nacu, 2010). Reads not uniquely mapping to the genome were removed, permitting default mismatches. Expression read counts of 62,069 ensembl genes were obtained by MMseq (Turro *et al.*, 2011).

### 3.2.3 Measuring DNA Methylation and Gene Expression

DNA methylation can be quantified by Beta-values (see formula 3.2) or M-values (see formula 3.3), where the Beta-value is the ratio of the methylated probe intensity and the overall intensity, and the M-value is the log2 ratio of the intensities of the methylated probe versus the unmethylated probe.

$$Beta_i = \frac{max(y_{i,methyl}, 0)}{max(y_{i,unmethyl}, 0) + max(y_{i,methyl}, 0) + \alpha} \tag{3.2}$$

$$M_i = log_2 \left( \frac{max(y_{i,methyl}, 0) + \alpha}{max(y_{i,unmethyl}, 0) + \alpha} \right) \tag{3.3}$$

$y_{i,methyl}$ and $y_{i,unmethyl}$ are the intensities measured by the $i^{\text{th}}$ methylated and unmethylated probe. The offset $\alpha$ is added to the denominator to regularize the value when both methylated and unmethylated intensities are low. The distribution of both Beta-values and M-values in our data can be seen in supplementary figure SF7 in Annex I.

The Beta-value results in a number between zero and one, which can also be interpreted as a percentage (0 or 100%). A value of one stands for complete methylation, a value of zero means that none of the measured molecules was methylated.

The M-value as a log ratio can be positive (more methylated) and negative (more unmethylated), where a value of zero represents intermediate methylation with half of the measured molecules methylated, and half of them unmethylated.

Beta-values can be transformed into M-values using formula 3.4 (Du *et al.*, 2010):

$$Beta_i = \frac{2^{M_i}}{2^{M_i} + 1}; M_i = log_2 \left( \frac{Beta_i}{1 - Beta_i} \right) \tag{3.4}$$

This conversion ignores the offset $\alpha$ of formula 3.2 and 3.3, which has been shown to only have negligible effects for the majority of probes (Du *et al.*, 2010).

We performed all analyses on M-values, except the part of the comparison of methods for measuring differential variability in DNA methylation data, where also the performance of Beta-values compared to M-values was investigated. Due to its easier interpretability as percentage of methylation however, we used the Beta-value when visualizing methylation values of genes exhibiting differential DNA methylation variability.

For the analysis of gene expression data, RNAseq counts (see section 3.2.2) were converted into expression log counts using R's function `log1p(x)`.

### 3.2.4 Analysis of DNA Methylation Variability

All the 423,089 probes that passed quality control and filtering procedures as described in section 3.2.2 were used in the analysis of DNA methylation variability.

For the comparison of different methods to measure differential variability we used the following R functions and packages:

- Bartlett's test (Bartlett, 1937) implemented in R's function `bartlett.test()`
- The Ansari-Bradley test (Ansari & Bradley, 1960) implemented in R's function `ansari.test()`
- Haim Bar's mixture model (Bar *et al.*, 2012) with code provided via personal communication (available at code_haim_bar.html) and calling the `harvest()` function
- DiffVar of the missMethyl package version 1.2.0 (Phipson & Oshlack, 2014) available from Bioconductor (2015)

Statistical significance was defined by Benjamini-Hochberg corrected p-values (Benjamini & Hochberg, 1995) smaller than 0.05.

The genomic annotation of the probes was made based on Illumina's manifest for the 450K BeadChip microarray (Illumina Inc, 2015). We defined "TSS200", "TSS1500", "5'UTR" and "1stExon" as belonging to gene promoters, and "Body" and "3'UTR" as belonging to gene bodies.

### 3.2.5 Analysis of Gene Expression Variability

For the analysis of gene expression variability, first of all we normalized the RNAseq dataset by library size using DESeq2 (Love *et al.*, 2014). Then we removed all genes with

no reads in more than 50% of the samples in one or more of the groups in order to only work with genes that are expressed in all three cell types. Furthermore, we included only protein coding genes. This led to a total number of 12,661 ensembl genes included in the analysis.

To measure differential expression variability with DiffVar (Phipson & Oshlack, 2014), the recommendation of its vignette was followed. In short, the count matrix was converted into a DGEList (Robinson *et al.*, 2009), a scaling normalization (Robinson *et al.*, 2009) was performed using the function `calcNormFactors()`, and a voom normalization was applied (Law *et al.*, 2014; Ritchie *et al.*, 2015; Phipson & Oshlack, 2014).

Statistical significance was defined by Benjamini-Hochberg corrected p-values (Benjamini & Hochberg, 1995) smaller than 0.05.

### 3.2.6 Analysis of Sex-Specific Differential Expression and DNA Methylation

Also for the analysis of gender differences within each cell type, the RNAseq data normalized by library size (see preceding section 3.2.5) was used. All genes without reads in more than 90% of the samples were removed, leading to a set of 38,824 genes. Genes of the sex chromosomes were maintained in the dataset to serve as a positive control, but excluded from any results reported. A scaling normalization (Robinson *et al.*, 2009) as well as voom normalization (Law *et al.*, 2014; Ritchie *et al.*, 2015) was performed before conducting the analysis of differential expression by limma (Smyth, 2005; Ritchie *et al.*, 2015). Genes were considered as significantly differentially expressed when their Benjamini-Hochberg corrected p-values (Benjamini & Hochberg, 1995) were smaller than 0.05.

Hypergeometric tests to assess whether overlaps between genes with increased gene expression variability and genes with sex-specific differential expression were bigger than expected by chance were performed using R's function `phyer()` on the basis of the 12,661 genes included in the analyses of differential variability.

Functional enrichment analyses of genes differentially expressed between males and females were performed with DAVID (Huang *et al.*, 2009) testing for GOTERM_BP_ALL, GOTERM_CC_ALL, GOTERM_MF_ALL. The same analysis was repeated using the R package GOseq (Young *et al.*, 2010). For both analyses the threshold of counts was set to a minimum of three genes and we considered terms as significantly enriched when the corresponding Benjamini-Hochberg adjusted p-values (Benjamini & Hochberg, 1995) were smaller than 0.05.

For the analysis of gender-specific differential methylation within each cell type we used the same limma model as for the analysis of sex-specific differential expression. No additional filtering was applied on the matrix of methylation M-values before performing the statistical analysis.

### 3.2.7 Programming Language

If not stated otherwise, the analyses were performed using R version 3 (R Development Core Team, 2008) and Bioconductor (Bioconductor, 2015).

# Chapter 4

# Results & Discussion

## 4.1 Gene Expression Variability in CLL

### 4.1.1 Overview

The focus of this work lies on the investigation of gene expression variability across individuals, an important parameter to be measured alongside the average levels of gene expression.

While studies on differences in mean levels of different human traits abound, few studies focused on variability so far, in part because of the larger numbers of observations necessary to accurately estimate variability which have only recently become available, and on the other hand because the importance of the information present in the distributions of biological data has often been overlooked so far. The gain of information from an improved understanding of phenotypic heterogeneity has only started to be appreciated in recent years (Ho *et al.*, 2008; Alemu *et al.*, 2014; Paszek *et al.*, 2010; Feinberg & Irizarry, 2010; Hansen *et al.*, 2011), as described in more detail in section 1.3. For CLL, there is increasing evidence that genetic and epigenetic heterogeneity are a crucial characteristic of disease development and progression (see section 1.5.2), but if and to which extent such heterogeneity is also present at the level of gene expression in CLL has not been previously investigated.

Here, we demonstrate that the two major clinical subtypes of CLL, M-CLL and U-CLL, which have very similar mean expression levels with only a small number of differentially expressed genes between the two subtypes (see also Ferreira *et al.* (2014) and Annex 2), show strong differences in gene expression variability, suggesting an impact of gene expression heterogeneity on tumor adaptability and aggressiveness in CLL.

The main results described in the following sections have also been published in Ecker *et al.* (2015).

### 4.1.2 Measuring Gene Expression Variability

In contrast to traditional analyses of differences in mean of any measured trait, here, we aimed to find genes with differential variability between groups, i.e. increased variability in one group compared to the other, as illustrated in figure 4.1, which shows the concept using synthetic data.

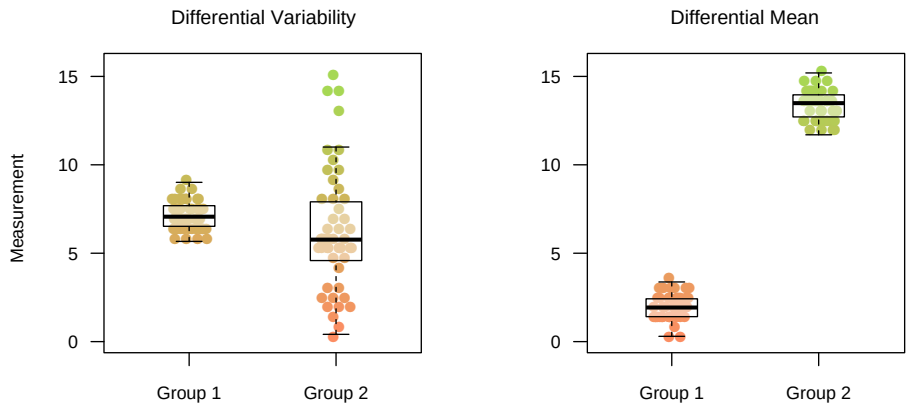


Figure 4.1: Differential variability versus differential mean in synthetic data. Data points represent individual measurements. These measurements could for example be gene expression values of a certain gene in all individuals within the group. In this case, every data point would represent the expression value of a gene in one individual, with multiple points representing different individuals of which the measurement of the same gene was obtained. Boxplots display the summary distribution of the underlying data points.

However, in real data, the two are not always as clearly separated from each other as in figure 4.1, and a gene can exhibit both, a significant difference in mean, but also in variability. It is important to take into account that there exists a dependence of expression variability on mean expression levels, making it sometimes hard to tear the effect of mean expression apart from variability measurements. Thus, we employed different measures of gene expression variability in order to obtain robust estimates of expression heterogeneity, and to avoid obtaining results of differential variability that are merely driven by differences in mean.

For the start, we studied gene expression variability in the ICGC dataset using the coefficient of variation (CV). The CV is defined as the ratio between the standard deviation (SD) of the variable measured across the patients and its mean (see section 3.1.2 and formula 3.1). As gene expression variability in terms of the CV is dependent on mean expression levels, we analyzed the dependence of the CV on the level of expression of the corresponding genes, see figure 4.2.

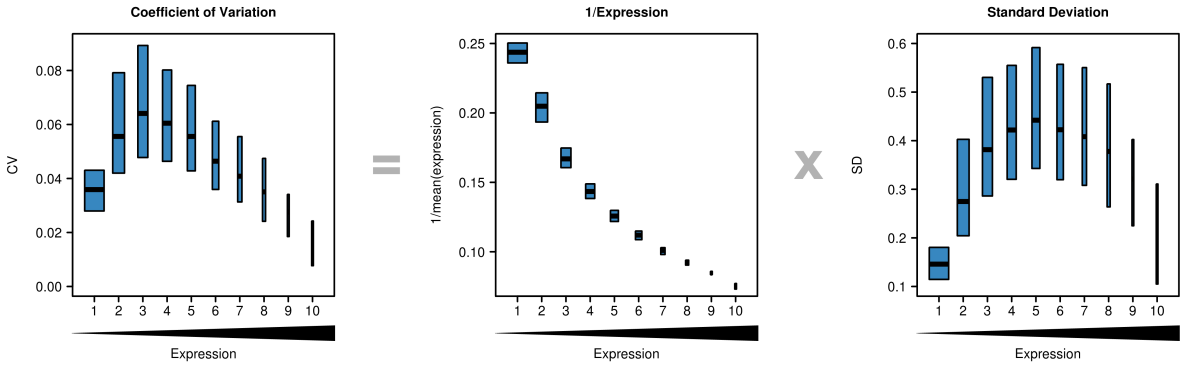


Figure 4.2: Definition of the CV and its dependence on gene expression levels. Left panel: CV versus expression of genes in bins of increasing expression level. Middle panel: Relationship between reciprocal of mean expression and expression in bins of increasing expression level. Right panel: Dependence of the standard deviation of expression across patients on the level of expression. The widths of the boxplots are relative to the number of genes contained in each gene expression bin (see also table 4.1).
Figure taken from Ecker *et al.* (2015).

The numbers of the genes contained in the bins shown in figure 4.2 are given in table 4.1. The relationship between the CV and mean expression levels is interesting and non-trivial. The highest levels of expression variability are observed for genes with low to intermediate levels of expression, and not for genes expressed at high or extremely low levels.

Table 4.1: Gene expression bins. The second column shows the expression values within every single bin and the last column lists how many genes are contained in the corresponding bin, also indicated by relative box widths in figure 4.2.

| Bin | Gene expression values | Nr of genes |
|---|---|---|
| 1 | $< 4.5$ | 9,101 |
| 2 | $\leq 4.5$ and $< 5.5$ | 3,355 |
| 3 | $\leq 5.5$ and $< 6.5$ | 2,402 |
| 4 | $\leq 6.5$ and $< 7.5$ | 1,987 |
| 5 | $\leq 7.5$ and $< 8.5$ | 1,517 |
| 6 | $\leq 8.5$ and $< 9.5$ | 986 |
| 7 | $\leq 9.5$ and $< 10.5$ | 435 |
| 8 | $\leq 10.5$ and $< 11.5$ | 224 |
| 9 | $\leq 11.5$ and $< 12.5$ | 74 |
| 10 | $\leq 12.5$ | 68 |

To understand the origin of this behavior it is important to take the intrinsic stochasticity of biological processes into account. The impact of fluctuations is inversely proportional to the number of elements involved in a system. This is a well-established phenomenon observed in physical systems (Kampen, 2007), and well characterized in biology (Kaern *et al.*, 2005; Lehner & Kaneko, 2011). Indeed, there is component of the CV that is given by the inverse of the mean of expression, as an $1/x$ dependence (see figure 4.2). This dependence reflects the fact that introducing an additional element in a small number

of observed traits – here, an extra copy of RNA of a lowly expressed gene – will have more dramatic consequences than an additional one in a huge amount of elements – here corresponding to an extra copy of a gene that is expressed in high numbers (see also figure 1.1). The latter will not produce a substantial change.

Certainly, stochastic processes of this kind are not likely to be the only determinants of expression variability. The remaining component of the CV is the SD, which has a negative quadratic dependence on the mean of expression (see figure 4.2), showing higher values for intermediate expression levels. Concluding, these observations highlight the importance of taking gene expression levels into account when evaluating expression variability.

Although the CV is known to be one of the most robust and unbiased metrics to quantify expression variability (Li *et al.*, 2010; Kaern *et al.*, 2005) and is the current gold-standard measurement, we employ an additional measure of expression variability which has recently been proposed by Alemu *et al.* (2014), subsequently called EV. The EV tries to account for the above described relationship between mean expression levels and variability in a distinct way and provides a measure of variability which is independent of the expression mean. It models variance as a function of the mean and gives the quantification of expression variability as the ratio of observed variance to expected variance for each gene (see section 3.1.2 and supplementary figure SF1 in Annex I for details on the method).

Despite the above described biological relevance of the relationship between mean expression levels and expression variability, introducing such a measure which "corrects" for the dependence of variability on mean expression levels has the advantage of being able to tear apart the effect of differing mean expression levels from variability estimations. Otherwise, increased expression variability observations in a gene could only be caused by differing mean expression levels.

We observed a high correlation between the CV and EV in the datasets we analyzed, with a Pearson correlation coefficient of 0.74 ($p < 2.2 \times 10^{-16}$) in the ICGC data and of 0.87 ($p < 2.2 \times 10^{-16}$) in the dataset of Fabris used for validation (the correlation can also be seen supplementary figure SF2 in Annex I). In all subsequent analyses we took both measures of gene expression variability into account.

### 4.1.3 Gene Expression Variability in the Two Subtypes of CLL

Next, we investigated whether gene expression variability differs between M-CLL and U-CLL, and could therefore be behind the different aggressiveness the two clinical dis-

ease subtypes. We plotted both the genes' CV and EV of M-CLL versus the CV and EV of U-CLL respectively, and consistent with our hypothesis gene expression variability shows a clear difference between the two subtypes, with higher variability associated to U-CLL, the more aggressive subtype of the disease (see figure 4.3).

As stated in Annex II and Ferreira *et al.* (2014) as well as other previous studies of gene expression in CLL (Klein *et al.*, 2001; Rosenwald *et al.*, 2001) and re-analyzed here using limma (Smyth, 2005; Ritchie *et al.*, 2015), mean gene expression values show only very little difference between the two types, see right panel of figure 4.3. Genes were considered differentially expressed when their Benjamini-Hochberg corrected p-values (Benjamini & Hochberg, 1995) were smaller than 0.05 and their absolute M-values were greater than 1.

These observations suggest that gene expression variability across patients can be an important factor to distinguish the two disease subtypes, for which differential mean expression is not discriminatory.

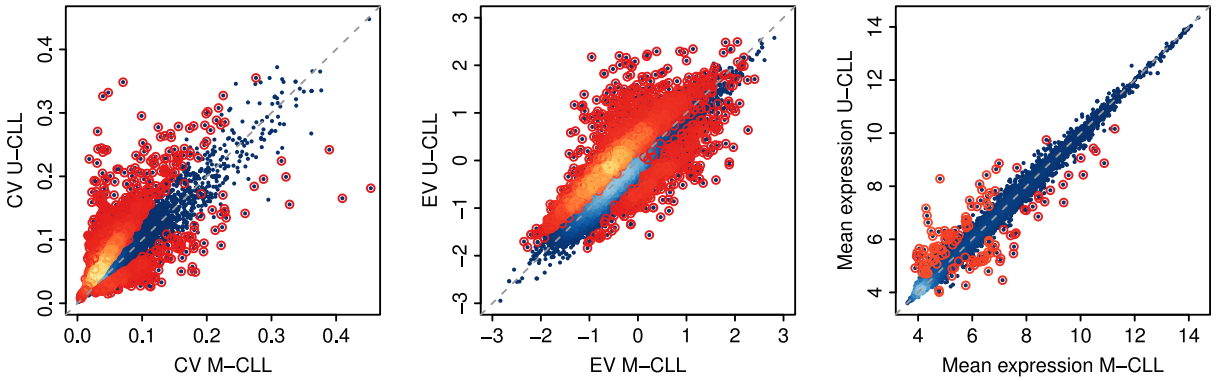


Figure 4.3: Gene expression variability comparison of U-CLL and M-CLL. Scatterplots comparing U-CLL and M-CLL where each data point represents a single gene. Lighter colors indicate higher densities of data points in the corresponding regions of the plot. Genes with statistically significant p-values at an false discovery rate (FDR) of 5% are highlighted by circles. The gray dashed line represents the identity line. Left panel: Scatterplot of CV across patients in the two disease subtypes. Genes with statistically significant differential variability according to the F-test are highlighted. Middle panel: Scatterplot of EV across patients in the two disease subtypes. Genes with statistically significant differential variability according to the F-test are highlighted. Right panel: Scatterplot of mean expression levels across patients in the two disease subtypes. Genes with statistically significant differential expression are highlighted.
Figure taken from Ecker *et al.* (2015).

As shown in figure 4.3, a substantial number of genes display higher variability across U-CLL patients compared to M-CLL patients. Applying an F-test (Snedecor & Cochran, 1989) with a FDR of 5% to assess statistical significance we found 2,025 genes with significantly increased variance in U-CLL whereas only 360 genes are significantly less variable in this subtype compared to M-CLL. Repeating these analyses with the smaller and less comprehensive datasets of Fabris and Haslinger used for validation, we confirmed

the result of increased expression variability in U-CLL, as can be seen in table 4.2 and supplementary figure SF3 in Annex I).

Table 4.2: Column 'all' lists the number of all genes, column 'sig' those with statistically significant p-values corresponding to the F-test (FDR=0.05). The significant genes are highlighted in figure 4.3 and SF3 correspondingly.

| | Increased var in M-CLL | | Increased var in U-CLL | |
|---|---|---|---|---|
| | all | sig | all | sig |
| ICGC | 6,425 | 360 | 13,871 | 2,025 |
| Fabris | 4,936 | 64 | 9,793 | 172 |
| Haslinger | 3,829 | 44 | 6,459 | 106 |

We found a strong correlation between the CV of the CLL subtypes in the patient cohorts of the ICGC and the data of Fabris (Pearson $r = 0.67$ in M-CLL and $r = 0.66$ in U-CLL, $p < 2.2 \times 10^{-16}$ in both, see also supplementary figure SF4 in Annex I), and also for the SD (Pearson $r = 0.75$, $p < 2.2 \times 10^{-16}$, supplementary figure SF4). The differences between the CV-values in M-CLL and U-CLL for genes in the two cohorts of the ICGC and Fabris are significantly correlated as well (Pearson $r = 0.28$, $p < 2.2 \times 10^{-16}$, supplementary figure SF4).

Furthermore, we observed a very high correlation of differential variability measured either by CV or EV differences (ICGC: Spearman correlation $\rho = 0.91$; Fabris: Spearman correlation $\rho = 0.93$; $p < 2.2 \times 10^{-16}$ in both; supplementary figure SF5 in Annex I). Comparing the top 500 genes with increased variability in U-CLL in each dataset (see supplementary table ST1 table_s1.html) we found a significantly higher than expected overlap (69 genes, hypergeometric test, $p < 2.2 \times 10^{-16}$).

Concluding, our results are reproducible in the two datasets investigated, both in terms of the correlation of the measurements of global expression variability of all genes, and in the comparison of ranked lists of the top differentially variable genes, and are therefore very unlikely to be caused by batch effects.

### 4.1.4 Gene Expression Variability and DNA Methylation

In the next step, we asked if the differences we observe in expression variability might be explained by differential DNA methylation. For the gene expression dataset of the ICGC matched DNA methylation data are available (Kulis *et al.*, 2012). Therefore, we compared the methylation profiles of the top 500 differentially variable (DV) genes with increased variability in U-CLL (supplementary table ST1 table_s1.html) but could not observe any strong and clear trend of different methylation levels between the two

subgroups M-CLL and U-CLL, not in the genes' promoters, neither in their bodies (see figure 4.4 and supplementary figure SF6 in Annex I).

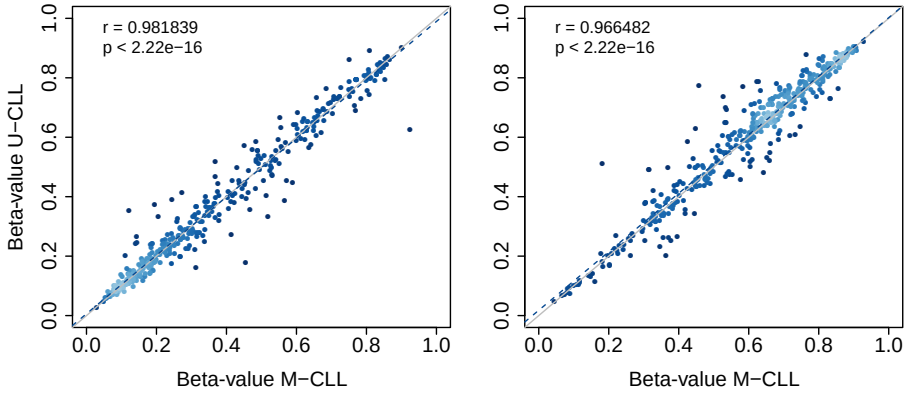


Figure 4.4: Methylation values of the top 500 genes with increased variability in U-CLL. Scatterplots comparing U-CLL and M-CLL. Lighter colors indicate higher densities of data points in the corresponding regions of the plot. The gray line represents the identity line, the blue dashed line shows a fitted regression line. Left panel: Promoter methylation. Right panel: Gene body methylation.
Figure taken from Ecker *et al.* (2015).

Additionally, we performed a region-based analysis of differential methylation between M-CLL and U-CLL (see section 3.1.3) in order to find out if methylation differences could relate to the differences in expression variability between the two subtypes. We identified 618 regions showing significant hypermethylation in U-CLL and 746 regions showing significant hypermethylation in U-CLL, but could again not find a direct relationship between DNA methylation and gene expression variability. Furthermore, neither the promoters nor the bodies of the top 500 DV genes with increased expression variability in U-CLL are represented within differentially methylated regions at higher rates than would be expected by chance, as can be seen in table 4.3.

Table 4.3: Results of hypergeometric tests assessing the overlap between genes within differentially methylated regions and genes with increased expression variability in U-CLL. The column 'all' lists the numbers of genes for which their promoters or gene bodies have been identified to lie within differentially methylated regions. Column 'DV' shows the number of the genes reported in the previous column which are also contained in the list of the top 500 genes with increased variability in U-CLL for which methylation measurements are available ($n = 491$). Column 'p' contains the p-values of the hypergeometric test evaluating if the overlap between the gene lists is bigger than expected by chance.

| | Hypermethylated in M-CLL | | | Hypermethylated in U-CLL | | |
|---|---|---|---|---|---|---|
| Region | all genes | DV genes | p-value | all genes | DV genes | p-value |
| Promoter | 381 | 15 | 0.2647 | 238 | 8 | 0.5174 |
| Gene body | 165 | 5 | 0.6296 | 173 | 6 | 0.4989 |

According to a study published by Lam *et al.* (2012), also DNA methylation variability in CLL does not have an obvious association with gene expression. Landau *et al.* (2014b)[1] however, employing a different approach to measure DNA methylation variability by cal-


[1]The study of Landau *et al.* (2014b) was published after our work (Ecker *et al.*, 2015) was accpeted for publication.

culating the proportion of discordant reads in regions of subsequent CpGs using whole genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) data, found that methylation variability in gene promoters was negatively correlated with average gene expression levels, and positively correlated with interindividual gene expression variability. However, as they stated in their study as well, it is difficult to say which of the two components is driving the correlation they identified, because of the strong negative dependence of expression variability on mean expression levels. Landau *et al.* (2014b) quantified gene expression variability using the CV and calculating the entropy of gene expression, without taking differing mean expression levels into account. Using an additional single-cell approach, they observed an association of disordered methylation with a decoupling of the expected relationship between promoter methylation and gene expression (described in section 1.2 and shown in figure 1.2).

Taking these results together, the relationship between DNA methylation and gene expression variability seems to be complex and difficult to disentangle. It is very likely that multiple layers of epigenetic modifications such as for example histone marks are involved in the process of regulating gene expression, or gene expression variability, and different dynamics of different epigenetic processes could contribute to this complexity. For example, it could also be possible that DNA methylation serves to poise genes for expression in response to future events (Lam *et al.*, 2012).

### 4.1.5 Functional Analysis of Differentially Variable Genes

We performed functional enrichment analysis on the top 500 genes with increased variability in U-CLL in order to detect if the increased variability in U-CLL affects specific biological processes. First of all, we identified the top 500 genes with increased variability in U-CLL in the ICGC dataset. We excluded all genes with non-significant p-values obtained by the F-test, and considered furthermore only genes with consistently increased variability in U-CLL across all three variability measures employed, that is, CV difference, EV difference, and the F-test (see section 3.1.2). The remaining genes were ordered by their CV differences and EV differences respectively. In the case of the dataset of Fabris only 172 genes reached statistical significance, therefore we did not apply the p-value cutoff here in order to achieve a comparable list of 500 genes. Both lists are available in supplementary table ST1 table_s1.html.

Functional enrichment analyses were then performed on these lists using both the webtool DAVID (Huang *et al.*, 2009) and Bioconductor packages (Falcon & Gentleman, 2007; Gentleman, 2015), see section 3.1.4 for further details.

Looking at the top 500 genes with increased variability in U-CLL patients of the ICGC study (see supplementary table ST1 table_s1.html) we observed a significant enrichment for processes related to the cell cycle, hemopoiesis, multicellular organismal processes, wounding, and development of the immune system and immune system processes. Altogether, we found 49 significantly enriched gene ontology (GO) terms and pathways at an FDR of 5% shown in supplementary table ST3 table_s3.html. Repeating the same analysis using the smaller dataset of Fabris, which interrogates less genes and comprises a smaller number of samples, we were able to recapitulate these findings to a certain extent, with significant enrichments in three immune system processes and, although not reaching statistical significance with a FDR of 5%, also in hemopoiesis, development, wounding, and cell proliferation (see supplementary table ST4 table_s4.html).

In order to gain a deeper understanding of the functional context of genes with significantly increased gene expression variability in U-CLL we performed network analyses. For these analyses we used the top 500 genes with increased variability in U-CLL in common in both datasets (ICGC and Fabris). We applied the same strategy as described above, with the only difference that we did not cut the list after the first 500 genes within each dataset separately but when reaching 500 genes in common in both datasets. The list of these top 500 genes is also available in supplementary table ST1 table_s1.html.

We mapped these top genes with increased variability in U-CLL to Entrez gene identifiers resulting in 494 unique Entrez genes in order be able to cross them with a B cell specific functional interaction network published by Lefebvre *et al.* (2010). This interactome assembles B cell specific transcriptional and post-translational molecular interactions that was constructed by Lefebvre *et al.* (2010) using a collection of 254 B cell gene expression profiles derived from normal and malignant B-cells of primary tumor samples and cell lines. The complete network contains 5,748 nodes (genes) and 64,600 unique edges (interactions). We extracted a subnetwork of the top genes with increased variability in U-CLL and included their direct neighbors into the subnetwork, considering only genes connected with at least two other genes. The resulting network contains 892 genes connected by 3,390 edges.

We identified five network modules in the network of 892 genes by using Gephi (Bastian *et al.*, 2009) and Louvain's method (Blondel *et al.*, 2008), an algorithm to detect communities in large networks based on modularity known to produce results of high quality (Newman, 2012; Lancichinetti & Fortunato, 2009) and which has been successfully applied in different network fields (Greene *et al.*, 2010; Zhang *et al.*, 2010; Meunier *et al.*, 2009; Amir *et al.*, 2013; Newman, 2012). Six genes were not mapped to any of the other

network modules and were therefore excluded from the subsequent functional enrichment analyses of network modules. Figure 4.5 shows the network with the five network modules highlighted in different colors.

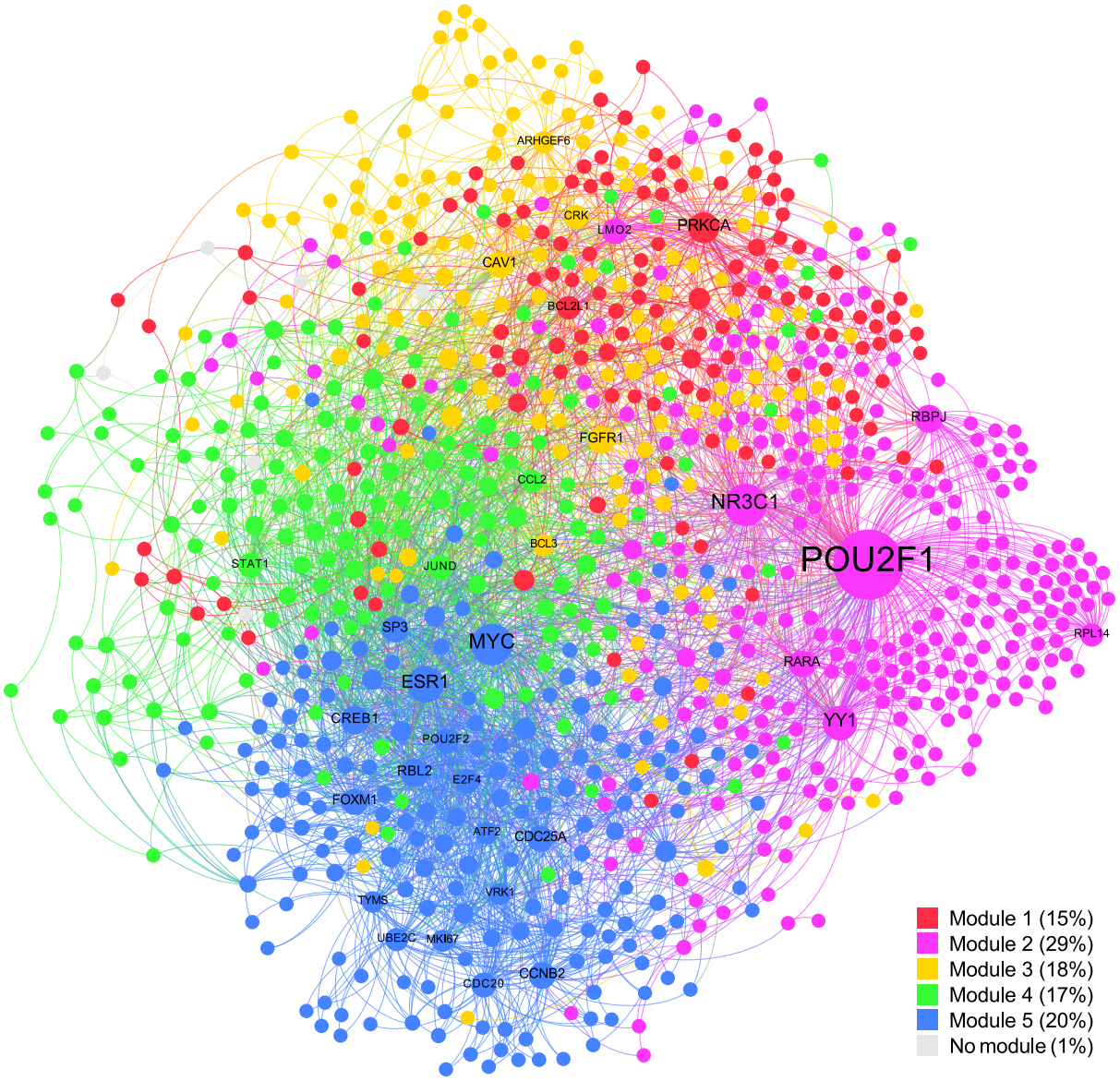


Figure 4.5: Network representation of genes with increased variability in U-CLL in the context of a B cell specific network [36]. Node sizes are determined by the degrees of the nodes, that is, big nodes represent highly connected genes. Different network modules are highlighted in different colors.
Figure taken from Ecker *et al.* (2015).

Functional enrichment analyses of the identified network modules showed that every module is highly enriched in biological processes and pathways, further confirming our results of biological functions affected by increased expression variability in U-CLL, and giving a deeper insight into these processes and pathways and the genes involved (see supplementary table ST5 table_s5.html and table 4.4 for a summary of the results).

Table 4.4: Functional enrichment of network modules. The first column shows the number of genes contained in every module of the network. The second column shows the top terms for which the corresponding module is enriched. The last column lists highly connected genes (degree $\geq 35$) of the corresponding module ordered alphabetically.

| | Genes | Top enriched terms | Highly connected genes |
|---|---|---|---|
| Module 1 | 135 | Cell death, cell differentiation and development | *BCL2L1, PRKCA* |
| Module 2 | 261 | Ribosome, translation | *LMO2, NR3C1, POU2F1, RARA, RBPJ, RPL14, YY1* |
| Module 3 | 160 | Signal transduction, cell communication, membrane, protein kinase activity, phosphorylation | *ARHGEF6, BCL3, CAV1, CRK, FGFR1* |
| Module 4 | 151 | Transcription factor activity, gene expression, DNA binding | *CCL2, JUND, STAT1* |
| Module 5 | 179 | Cell cycle | *ATF2, CCNB2, CDC20, CDC25A, CREB1, E2F4, ESR1, FOXM1, MIKI67, MYC, POU2F2, RBL2, SP3, TYMS, UBE2C, VRK1* |

The first network module is heavily enriched for cell death and apoptosis, and shows furthermore enrichments for cell differentiation, cellular development processes, and system and multicellular organismal development as well as cancer pathways. The most connected gene in this module is *PRKCA* (protein kinase C alpha), a kinase involved in cell differentiation, cell cycle checkpoint and cell volume control which also plays an important role in the growth and invasion of cancers (Koivunen *et al.*, 2006) and is known to act as an anti-apoptotic agent in leukemic B cells by phosphorylating *BCL2* (Ruvolo *et al.*, 1998). Precisely *BCL2L1* (B cell CLL/lymphoma 2 like 1), a member of the *BCL2* family, is the second most connected gene in the module, and has also been suggested to play an important role in B cell apoptosis and CLL (Jiang & Clark, 2001). Two alternatively spliced transcript isoforms of the gene are currently known, where the longer isoform acts as an apoptotic inhibitor and the shorter one as an apoptotic activator (Boise *et al.*, 1993).

Module two of the network, which is enriched for the ribosome and translation as well as transcription, contains the biggest hub of the network, *POU2F1* (POU class 2 homeobox 1), a transcription factor which has been associated with the cell cycle (Roberts *et al.*, 1991; Segil *et al.*, 1991) and is involved in the activation of immunoglobulin genes (Lee *et al.*, 2001). *POU2F1* has also been related to the deletions on chromosome 11 in CLL (Auer *et al.*, 2005).

Further highly connected genes in module two are the glucocorticoid receptor *NR3C1* (nuclear receptor subfamily 3, group C, member 1) which regulates developmental genes and affects inflammatory response, cellular proliferation and differentiation, and has also been related to the cell cycle (Lu *et al.*, 2006), and *YY1* (yin yang 1), a transcription factor involved in the activation and transcription of ribosomal proteins (Voronina *et al.*, 2008), development and differentiation, as well as tumorigenesis (Sui, 2009). The gene has been related to a plethora of human cancers and hematopoietic malignancies (Bonavida *et al.*, 2011; Nicholson *et al.*, 2011).

Other smaller hubs in module two are the gene *LMO2* (LIM domain only 2 rhombotin-like 1), an oncogene which plays a crucial role in hematopoietic development and leukemia (Warren *et al.*, 1994; Davenport *et al.*, 2000), and *RPL14* (ribosomal protein L14) and *RARA* (retinoic acid receptor alpha), both associated with translation (Odintsova *et al.*, 2003; Chen *et al.*, 2008).

The signaling module (module three) in the network shows – beside heavy enrichments for signal transduction and cell communication – localization to the plasma membrane and further enrichments for kinase activity and phosphorylation. One of the highly connected genes within this module is *CAV1* (caveolin 1), a gene strongly related to signal transduction which is able to affect cell function and cell fate (Shatz & Liscovitch, 2004; Engelman *et al.*, 1998) and has furthermore been described to play a significant role in CLL progression (Gilling *et al.*, 2012).

Beside further important signaling genes like *FGFR1* (fibroblast growth factor receptor 1), a cell-surface receptor playing an essential role in development, cell proliferation, differentiation and migration (Groth & Lardelli, 2002) which has also been related to clinical outcome in CLL (Gilling *et al.*, 2012), another highly connected gene in network module three is *BCL3* (B cell CLL/lymphoma 3) which is involved in the activation of *NF-κB* target genes, plays a role in the regulation of cell proliferation (Na *et al.*, 1999), and has been described to stimulate *AP1* (activator protein 1) proteins (Na *et al.*, 1999).

In the context of signaling in CLL, it is important to mention that the BCR – also contained in network module three, although not among the most highly connected genes in the module – has been reported to exhibit crucial differences in M-CLL and U-CLL. As also described in section 1.5 and Annex II, BCR signaling leads to transcriptional responses in CLL cells that have been strongly associated with cell activation, enhanced cell cycle entry, and progression of the disease. After stimulation of the BCR in CLL, key molecules of the BCR signaling cascade (for example *ZAP70* and *SYK*) are recruited,

which leads to the phosphorylation of the B cell linker protein, a central node for intracellular signaling (Zenz *et al.*, 2010).

In U-CLL, a higher proportion of stereotyped rearrangements and biased somatic mutation patterns could be observed in the BCR (Messmer *et al.*, 2012), and the BCR in U-CLL has been reported to be usually polyreactive to autoantigens, for example proteins or lipids generated by oxidative stress (Chiorazzi & Ferrarini, 2003, 2011). It was concluded that in U-CLL the BCR signaling pathway is more readily stimulated by cross-linking (Rosenwald *et al.*, 2001; Zenz *et al.*, 2010; Guarini *et al.*, 2008). Indeed, several genes have been shown to be exclusively modulated in U-CLL cells upon BCR activation, indicating that antigenic stimulation plays a key role in the progression of CLL, as the BCR signaling pathway is associated with cell-cycle progression and survival of malignant B cells (Chiorazzi & Ferrarini, 2011; Guarini *et al.*, 2008).

However, there is some controversy about the differences of the BCR in the two disease subtypes in the literature. Herishanu *et al.* (2011) for example showed evidence of BCR activation in both U-CLL and M-CLL cells *in vivo*, at least in the lymph node environment. Similar results have been reported by Krysov *et al.* (2012) and Pede *et al.* (2013), who found both M-CLL and U-CLL cells to respond to BCR activation, and concluded that the difference observed in freshly isolated peripheral blood CLL cells is caused by differences in *in vivo* triggering of the BCR (Pede *et al.*, 2013), consistent with the observation of Herishanu *et al.* (2011) that the transcriptional differences are small in cells isolated from lymph nodes.

Differences between M-CLL and U-CLL in the response of the BCR could lead to a bias due to isolation procedures in *in vitro* experiments, as positive antibody selection activates signaling pathways and induces gene expression changes not inherent to the cell but caused by the activation of the BCR. In any case, the documented relationship between expression variability and adaptability in response to perturbations (see section 1.3) suggests that the higher heterogeneity present in this disease subtype might contribute to the increased response in signaling reported for U-CLL.

Interestingly, in the ICGC study of Ferreira *et al.* (2014) in which we investigated the transcriptome of CLL (see also Annex II), splicing changes were identified in several genes of the BCR pathway. These changes could possibly contribute to the observed variability in signaling in U-CLL, as there is evidence of an association between alternative splicing and gene expression variability (Wang & Zhou, 2014).

Signaling has also been reported to be deregulated in CLL and especially in U-CLL by others (Chuang *et al.*, 2012; Guarini *et al.*, 2008; Landau *et al.*, 2013), and signaling pathways were therefore suggested as potential targets for treatment strategies (Zenz *et al.*, 2010; Kipps, 2007).

In network module four, which is enriched for transcription factor activity, DNA binding and gene expression, the most connected gene is *JUND* (jun D proto-oncogene), a member of the above mentioned *AP1* transcription factor complex that regulates lymphocyte proliferation (Meixner *et al.*, 2004). It has been suggested to protect cells from *P53* induced senescence and apoptosis (Weitzman *et al.*, 2000) and has an influence on tumorigenesis and cancer progression (Eferl & Wagner, 2003).

Two other highly connected genes of network module four are *CLL2* (chemokine C-C motif ligand 2), a gene involved in immunoregulatory and inflammatory processes (Xu *et al.*, 1996) which has *AP1* binding sites in its promoter (Wolter *et al.*, 2008), and *STAT1* (signal transducer and activator of transcription 1, 91kDa), a transcriptional activator which plays an important role in lymphocyte proliferation and survival as well as cell viability in response to stimuli and pathogens (Lee *et al.*, 2000). In CLL *STAT1* has furthermore been shown to be related to resistance to DNA-induced apoptosis (Vallat *et al.*, 2003) and to be aberrantly phosphorylated on serine residues (Frank *et al.*, 1997).

The most important gene of the cell cycle module (module number five) is *MYC* (v-myc avian myelocytomatosis viral oncogene homolog), a transcription factor that activates the expression of many genes but has also been suggested to act as a transcriptional repressor (Pelengaris *et al.*, 2002). It is a key regulator of cell cycle entry (Krysov *et al.*, 2012), has a direct role in the control of DNA replication (Dominguez-Sola *et al.*, 2007), and regulates differentiation, cell growth and apoptosis by modulating the expression of distinct target genes like for example the downregulation of *BCL2* among other apoptotic pathway genes (Pelengaris *et al.*, 2002; Lüscher, 2001; Nilsson & Cleveland, 2003). Deregulation of *MYC* has been shown to be very strongly related to tumor formation (Lüscher, 2001), and the expression of *MYC* is altered in many types of cancer (Nilsson & Cleveland, 2003), including CLL (Rana *et al.*, 2014), where it has been demonstrated that *MYC* and its target genes are overexpressed in lymph nodes compared to blood cells (Herishanu *et al.*, 2011), and that increased basal expression of *MYC* in CLL cells is associated with progressive disease (Zhang *et al.*, 2010).

Further highly connected genes in the cell cycle module are *FOXM1* (forkhead box protein M1) which plays a key role in multiple facettes of cell cycle progression and is known as

a proto-oncogene which contributes to both tumor initiation and progression in leukemia (Wierstra & Alves, 2007; Mencalha *et al.*, 2012), and has been shown to be upregulated in many tumors, and other key regulators of the cell cycle such as *ESR1* (estrogen receptor 1) which is known to be involved in cell growth, cellular proliferation and differentiation (Shupnik, 2004), *RBL2* (retinoblastoma-like 2), a progression marker gene in CLL (Fält *et al.*, 2005), and *E2F4* (E2F transcription factor 4, p107/p130-binding), a gene which has been shown to be deregulated in rapidly growing B cell lymphomas. The latter two are also interacting key regulators of the cell division cycle (Sardet *et al.*, 1995). Beside many other cell cycle related genes highly connected in network module five it also contains the gene *MKI67* (marker of proliferation Ki-67), a widely used marker of cellular proliferation in human tumors and a strong predictor of survival in CLL (Bruey *et al.*, 2010).

The heavy enrichment for cell cycle related genes and functions among the genes with increased variability in U-CLL is especially interesting because historically, CLL was viewed as a purely accumulative disease of malignant cells with a defect in apoptosis (Rosenwald *et al.*, 2001; Rozman & Montserrat, 1995; Caligaris-Cappio & Hamblin, 1999). Recent studies however showed that proliferation plays an important role in CLL progression (Messmer *et al.*, 2005; Guarini *et al.*, 2008; Chiorazzi *et al.*, 2005; Obermann *et al.*, 2007; Krysov *et al.*, 2012). A study of Messmer *et al.* (2005) measuring cell kinetics using a nonradioactive method showed for example that the above mentioned cell cycle marker *MKI67* is expressed in CLL *in vivo*.

Obermann *et al.* (2007) showed two years later that although the majority of CLL cells are resting in G0 phase, a considerable number of cells have proliferative potential, with a significant subpopulation of cells residing in early G1 phase, probably contributing to a more aggressive biological behavior with increasing numbers. In their study the amount of cells with proliferative potential indicated by the expression of the cell proliferation marker *MCM2* (minichromosome maintenance complex component 2) – a key component in genome replication (Tye, 1999; Chong *et al.*, 1996)) – was far higher than the number of proliferating cells expressing *MKI67*. The progression into the G1 phase is particularly important as cells in the G1 phase are known to be more prone to external stimuli to further progress into cell cycle (Obermann *et al.*, 2007), which relates again to the increased variability observed here, and to the significant enrichment of signaling in the network of highly variable genes in U-CLL.

Again, the BCR seems to play an important role in this context. BCR stimulation via antigens has been demonstrated to induce G1 progression, proliferation and enhanced

cell survival in U-CLL patients (Longo *et al.*, 2007; Deglesne *et al.*, 2006). Several studies found specific expression of proliferative pathways in U-CLL (Kanduri *et al.*, 2010; Guarini *et al.*, 2008; Herishanu *et al.*, 2011), while M-CLL showed an increase of apoptotic levels in stimulated cells (Guarini *et al.*, 2008). Related to these observations, Messmer *et al.* (2005) reported a wide range of different cell proliferation and apoptosis rates across CLL patients, further supporting the findings of increased variability in the corresponding network modules analyzed here.

Consistent with the observation of increased proliferative potential in U-CLL, it has moreover been shown that the telomeres of U-CLL cells are much shorter than those of cells obtained from age-matched normal donors, but also compared to M-CLL cases (Hoxha *et al.*, 2014; Sellmann *et al.*, 2011; Damle *et al.*, 2004; Rampazzo *et al.*, 2012), indicating an extensive history of cell division, and pointing to the activation of pathways inducing proliferation in U-CLL cells. Telomere shortening is also contributing to genomic instability and could enhance the clonal evolution predominantly observed in U-CLL (see section 1.5.2).

Furthermore, Rana *et al.* (2014) reported a deregulation of circadian clock genes in CLL. The aberrant expression of these genes could contribute to genomic instability and accelerated proliferation due to the aberrant expression of downstream targets involved in cell proliferation and apoptosis (Rana *et al.*, 2014).

Some authors even suggested that proliferation and not apoptosis inhibition might be the main criterion determining clinical outcome in CLL, and cell cycle inhibitors have already entered clinical trials as therapeutic agents (Herishanu *et al.*, 2011; Obermann *et al.*, 2007; Flynn *et al.*, 2015).

Concluding, U-CLL patients show increased variability in cell proliferation directly affected by key cell cycle regulation genes, and furthermore in cell differentiation and development, cell death, and intercellular communication and signaling, all of which could be impacting the aggressiveness and adaptability of this subtype, possibly explaining the worse clinical outcome of U-CLL.

### 4.1.6 Classification of Patients by Gene Expression Variability

The previously described results showing considerable differential gene expression variability between the two subtypes of CLL suggest that measurements of gene expression heterogeneity might be a distinctive feature of M-CLL and U-CLL that can be used for the separation of the two subtypes in a classification approach.

As we have stated previously here and in Ferreira *et al.* (2014) and had also been reported by others before (Klein *et al.*, 2001; Rosenwald *et al.*, 2001), gene expression data "as is" is not sufficient to cluster patients into the two disease subtypes M-CLL and U-CLL. For example, considering the gene expression levels of all genes from the ICGC gene expression dataset ($n = 20,149$) and applying a standard hierarchical clustering, no separation of the two disease subtypes is obtained, as can be seen in figure 4.6.

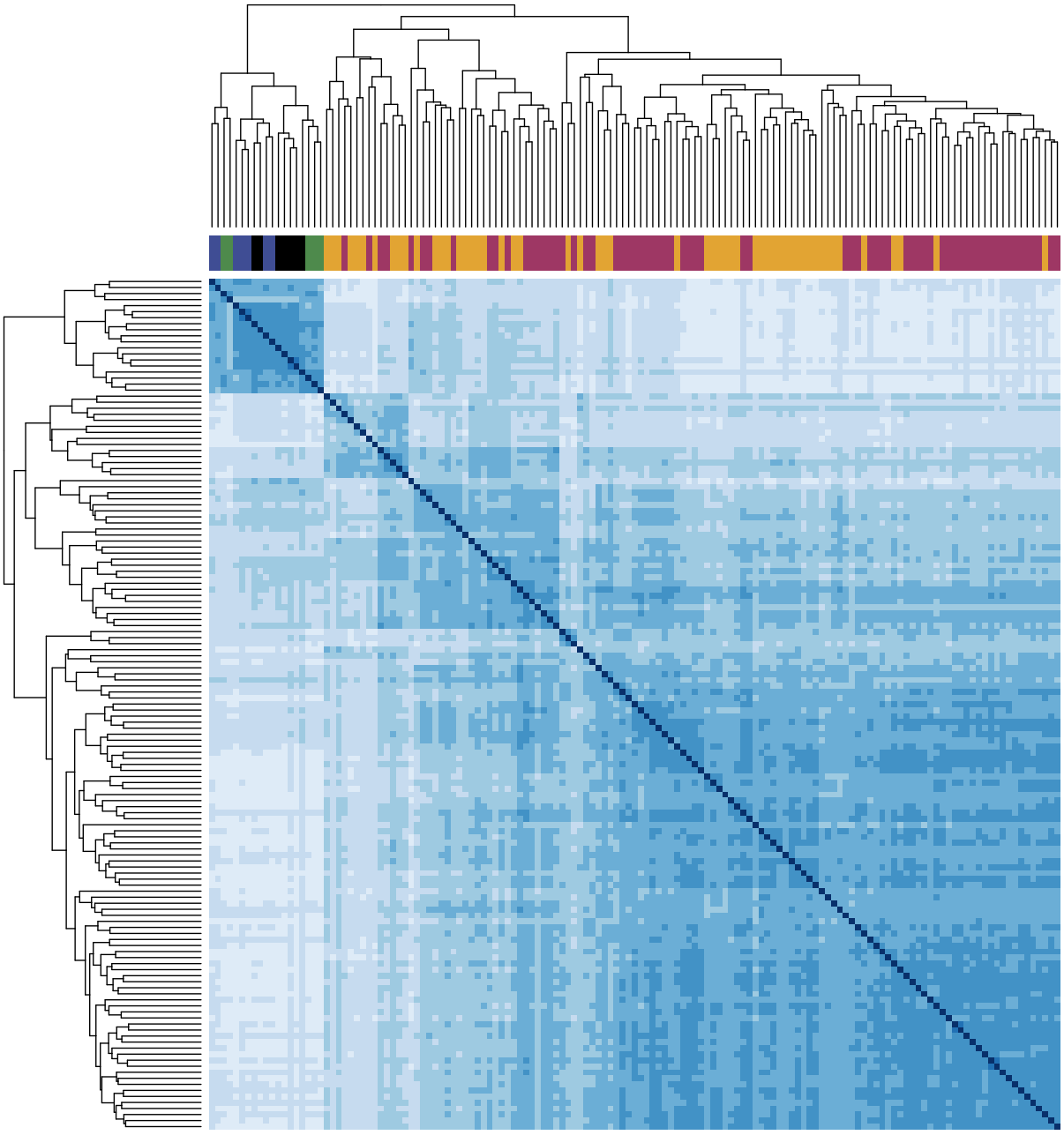

Figure 4.6: Hierarchical clustering of gene expression data. Heatmap representing a clustering of the CLL samples of the ICGC study. Dark blue colors in the heatmap represent short distances, light colors indicate large distances. U-CLL samples are colored in orange, M-CLL samples in dark magenta, and healthy cells in blue (memory B cells), green (naive B cells) and black ($CD19^+$ B cells).
Figure taken from Ecker *et al.* (2015).

Nevertheless, applying a kind of "de-noising" strategy on the expression data allowed us to group the patients reasonably well into the two subtypes via unsupervised clustering. This approach works by aggregating patients into groups by extracting five random U-CLL patients and another five random M-CLL patients as long as sets of five can be made without repeating samples in the groups, such that a new cohort of "superpatients" is produced, in which half are M-CLL and the other half U-CLL superpatients. These superpatients represent constructs of aggregated samples that help to remove noise from the data. Indeed, when we calculated the mean expression values for the superpatients, we were now able to separate the two groups by hierarchical clustering, as can be seen in figure 4.7.

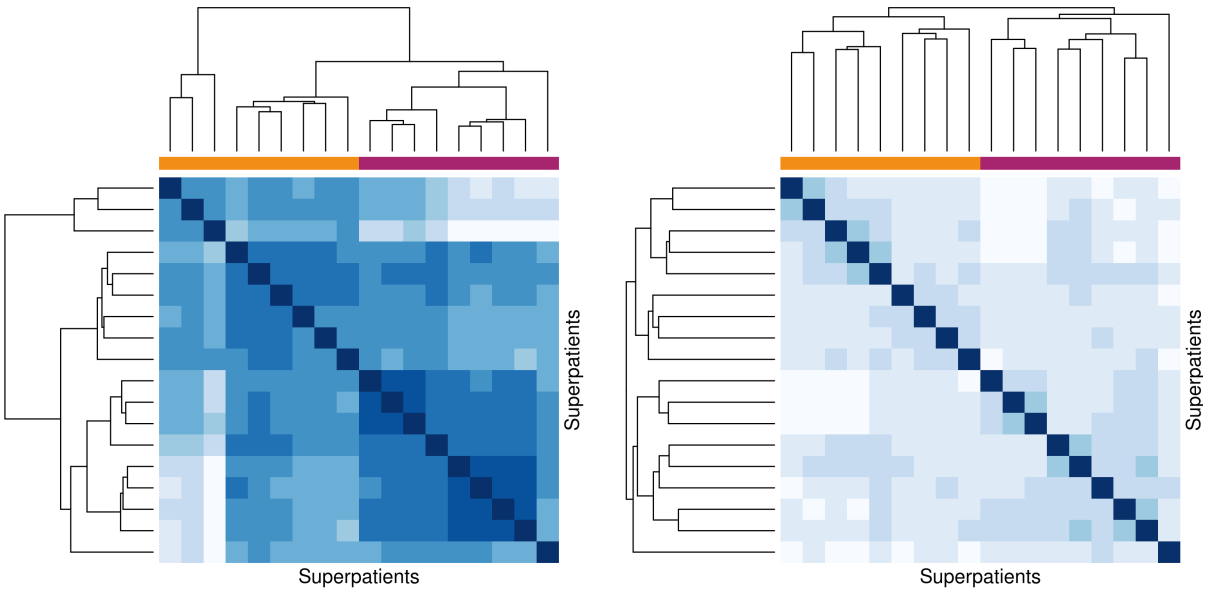


Figure 4.7: Hierarchical clustering of superpatients. Heatmaps representing the clustering of superpatients into M-CLL and U-CLL, based on aggregate measures. Left panel: Superpatient clustering based on mean expression values. Right panel: Superpatient clustering based on the CV. Results very similar to the figure on the right were obtained when using other measurements of variability such as the EV, SD, interquartile range (IQR) or different distance measures (data not shown).
Figure taken from Ecker *et al.* (2015).

This approach provides the additional advantage that also variability of gene expression can be measured, which is not possible for individual patients when not having multiple samples of the same patient across different time points. A hierarchical clustering based on variability measurements also separates the two subgroups very well (see figure 4.7).

Changing the number of patients used to create the superpatients like for example taking seven or ten random samples instead of five did not generally alter the results. However, as the superpatient approach relies on random subsampling of the patients, the results of the hierarchical clusterings can vary after every new run of patient aggregation. This

approach should therefore not primarily be taken as a stable classification method but more as a demonstration of the power of the concept. The fact of being able to separate the two disease subtypes by applying the superpatient method indicates that the previous observations of gene expression profiles not being able to distinguish the two disease subtypes are probably caused by noise and variation of both technical (Tu *et al.*, 2002; Gentleman *et al.*, 2005) and biological (as described in this work) origin which is present in transcriptomic data.

To further investigate this observation, we sought to apply an unbiased classification method and trained a random forest classifier (Breiman, 2001; Liaw & Wiener, 2002) on the ICGC gene expression data using 1,000 trees. Subsequently, we used this classifier to predict the CLL subtypes of the patients in the dataset of Fabris. To establish the feature sets used in the random forest classification approach we considered only genes present on both microarray platforms of the two different studies ($n$ = 12,307). In order to robustly estimate error rates, we repeated the analysis 1,000 times.

First, mean expression values of the 12,307 genes present in both datasets were used as features. The resulting random forest classifier based on gene expression values was able to classify patients correctly, with a mean AUC of 0.90, see figure 4.8 and table 4.5.

Next, based on the observations we made when reducing gene expression noise (see above), we repeated this analysis using only the top 500 genes with most different mean expression levels between M-CLL and U-CLL corresponding to their absolute M-values. The list of these genes can be found in the first column of supplementary table ST2 table_s2.html. The prediction of the disease subtypes in Fabris' dataset based on the classifier trained on the ICGC data improves considerably when using the top 500 differentially expressed genes, now reaching a mean AUC of 0.96 (see figure 4.8 and table 4.5).

Finally, inspired by our promising results on the importance of the variability of gene expression as a defining characteristic of M-CLL and U-CLL, we created random forests using the top 500 most differentially variable genes. Here, we only included genes with Benjamini-Hochberg corrected p-values (Benjamini & Hochberg, 1995) smaller than 0.05 and furthermore we only took genes into account which showed consistently increased or decreased variability according to all three differential variability measures we applied (that is, the CV difference, the EV difference, and the F-test) and ordered the list of the remaining genes once by their absolute CV differences, and once by their absolute EV differences. The top 500 most differentially variable genes corresponding to the CV and EV are available in column two and three of supplementary table ST2 table_s2.html).

We defined a new feature to measure expression variability for each gene in each patient as the distance from a gene's expression value $x_i$ to the median of that gene $i$ over the population $\tilde{x}_i$, see formula 4.1.

$$dist_i = |x_i - \tilde{x}_i| \tag{4.1}$$

We trained our random forest classifier applying this measure to the top 500 differentially variable genes on the data of the ICGC aiming again to predict the disease subtype of the patients in Fabris' dataset. Strikingly, this classifier based on gene expression variability performs equally well as the one based on differential expression, with a mean AUC of 0.96, and an even smaller standard deviation thereof, indicating more robust results compared to using mean expression levels (see figure 4.8 and table 4.5).

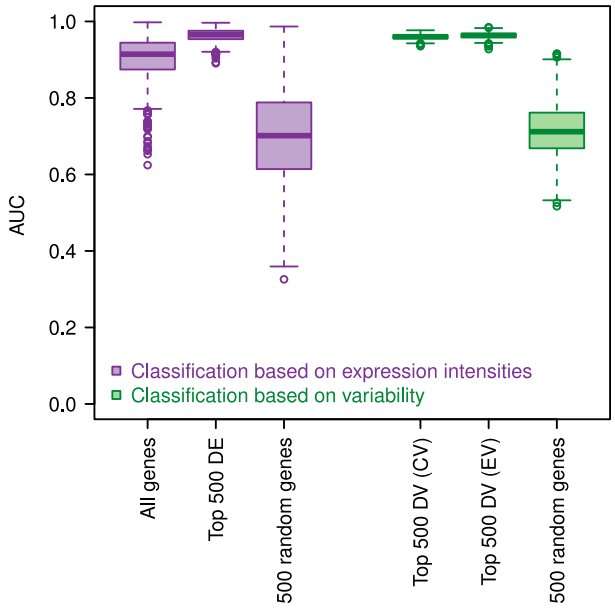


Figure 4.8: Random forest classifier results. Boxplots showing the distribution of AUC values of 1,000 independent runs per classifier.

Figure taken from Ecker *et al.* (2015).

Classifiers using feature sets consisting of 500 randomly selected genes perform significantly worse, both in the case of using mean gene expression levels as well as when using the variability measure introduced above. The results of the different classifiers are shown in figure 4.8 and table 4.5.

Table 4.5: Random forest classifier results. AUC values of 1,000 independent runs per classifier. The first three rows show the results of the classifier when it is feeded with gene expression values (exprs values). The last three rows show the result of the classifier when the features are selected based on expression variability, and gene-wise variability was quantified by the above introduced variability measure which calculates the absolute distance of a gene's expression value from the population median (var values).

| | Mean | Median | Min | Max | SD |
|---|---|---|---|---|---|
| All genes (exprs values) | 0.9028 | 0.9144 | 0.6244 | 0.9977 | 0.0578 |
| Top 500 DE genes (exprs values) | 0.9637 | 0.9653 | 0.8906 | 0.9965 | 0.0162 |
| 500 random genes (exprs values) | 0.7000 | 0.7014 | 0.3258 | 0.9867 | 0.1218 |
| Top 500 DV genes based on CV (var values) | 0.9596 | 0.9601 | 0.9352 | 0.9769 | 0.0064 |
| Top 500 DV genes based on EV (var values) | 0.9635 | 0.9635 | 0.9277 | 0.9850 | 0.0079 |
| 500 random genes (var values) | 0.7172 | 0.7118 | 0.5168 | 0.9161 | 0.0736 |

In summary, our initial results on the available datasets suggest that expression variability can classify the two clinical subtypes of CLL very well, pointing to a potential relation between expression variability and disease aggressiveness. If these results can be confirmed by follow-up studies with larger datasets, it will be interesting to explore the use of expression variability for the classification of other disease states as well.

### 4.1.7 Interpretation and Further Discussion

As has already been reported for heterogeneity in CLL at the genetic and epigenetic level (Kleppe & Levine, 2014; Swanton & Beck, 2014; Oakes *et al.*, 2014; Landau *et al.*, 2013, 2014b), increased variability in CLL worsens clinical outcome, perfectly in line with our findings of increased gene expression variability in the more aggressive type of the disease, which had not been investigated before. Taking these observations into a more clinical point of view, monitoring gene expression heterogeneity during disease course might be an advantageous strategy to improve therapeutic decisions and risk prediction in CLL patients. Additionally, the further assessment of the implications of variability in CLL might gather important insights into the disease and provide knowledge for new combinatorial therapies (Landau *et al.*, 2014a).

The patients included in this study did not receive any prior CLL treatment, but it has to be noted that we cannot exclude the possibility of (other) drug therapy that U-CLL patients might have received, the age of the individuals, or eventual technical factors to contribute to the observed increase in interpatient variability in U-CLL.

What can however be excluded as a confounding factor in this work are cell type composition effects, as CLL tumor samples can be obtained at near-complete purity with $\geq 95\%$ neoplastic cells (Kulis *et al.*, 2012; Ferreira *et al.*, 2014). Additionally, CLL genomes are nearly diploid and gene expression variability is therefore unlikely to be influenced by

somatic copy number variation (Landau *et al.*, 2014b), while other genetic and epigenetic alterations, as described previously (see section 1.3 and 1.5.2), probably play a role together with gene expression variability in generating phenotypic diversity in CLL.

Interestingly however, genetic heterogeneity in terms of higher frequencies of somatic mutations has been reported to be predominantly associated with M-CLL, not U-CLL (Landau *et al.*, 2013; Puente *et al.*, 2011; Quesada *et al.*, 2011), while clonal evolution and increased methylation heterogeneity have been associated to U-CLL (Oakes *et al.*, 2014; Stilgenbauer *et al.*, 2007; Herishanu *et al.*, 2011; Landau *et al.*, 2014a; Gunnarsson *et al.*, 2011), the subtype of the disease in which we observe increased gene expression variability in our work as well.

Another notion that has to be made is that for human data not much is yet known about the translation of cellular RNA expression variability into protein levels (Satija & Shalek, 2014), although several studies tried to elaborate on the question of the relationship between gene expression variability and protein fluctuations in model systems (Ozbudak *et al.*, 2002; Blake *et al.*, 2003; Kaern *et al.*, 2005; Raj & Van Oudenaarden, 2008; Golding *et al.*, 2005; Kierzek *et al.*, 2001; McAdams & Arkin, 1997).

Finally, our observation of increased expression variability across patients in U-CLL is likely to relate to single cell heterogeneity within patients as explained in section 1.3. Actually, relying on this hypothesis, we can link the variability observed across individuals to the worse prognosis for U-CLL patients, which can be attributed to the presence of cellular heterogeneity and hence aggressiveness, adaptability, and increased resistance to therapy in this disease subtype. However, to verify this hypothesis, larger datasets comprising multiple samples per patient and/or time series, and in particular single cell data allowing the detailed study of transcriptional variability will be an invaluable new source of complementary information.

The here presented work is the first systematic exploration of gene expression variability in CLL. Our results provide important additional biological insight into the mechanisms of the disease and uncover previously unappreciated dynamics present in CLL. Based on these findings it might be possible in the future to develop better prognostic tools and new therapeutic strategies improving patient outcome in CLL.

## 4.2 Variability in Normal Blood Cells

### 4.2.1 Overview

Blood cells, as explained in section 1.3.2, exhibit a remarkable variability in terms of size and characteristics. If two sister cells are separated directly after cell division and get cultured under identical conditions, they can build colonies of different blood cell types, or of the same cell type, but in distinct amounts (Alberts *et al.*, 2004). In the programming of cell division as well as in the process of determining the differentiation branch followed, and in the gradual regulation of the multicellular system as a whole in response to changes in the environment, stochastic processes at the level of individual cells are involved (Hume, 2000; Satija & Shalek, 2014; Enver *et al.*, 1998; Chang *et al.*, 2008). To fully understand these processes and the interplay between different layers of gene regulation as well as their heterogeneity they need to be analyzed in a combined way.

We used a dataset of the Human Variation Epigenome Project of BLUEPRINT (Adams *et al.*, 2012) to analyze variability in DNA methylation and gene expression data of monocytes, neutrophils and T cells derived from 48 healthy individuals (see also section 3.2.1). The dataset of this project will be extended to 200 individuals in the future. An overview of the data can be seen in figure 3.1. The pilot dataset of 48 individuals was used to develop the methodology and analysis pipelines for the BLUEPRINT project and to provide hypothesis to be tested specifically later in the project with the larger dataset.

In the following we discuss how to measure and compare variability across different cell types in distinct data types and present initial results obtained from the analysis of the pilot dataset, showing that neutrophils exhibit patterns of increased interindividual variability compared to monocytes and T cells.

### 4.2.2 Analysis of Variability in Different Biological Data Types

As here we want to analyze both DNA methylation microarray data and RNAseq data to investigate DNA methylation and gene expression variability as well as their interrelationships, we needed to establish a methodological framework that is able to deal with both microarray and sequencing data and to produce comparable results across distinct biological data types.

To this aim, we used DiffVar (Phipson & Oshlack, 2014), a recently developed method to measure differential variability based on limma (Smyth, 2005), a sophisticated R software package originally developed for the analysis of differential gene expression in microarrays

that has been continuously extended during the past decade and is now also able to deal with RNAseq and DNA methylation data (Ritchie *et al.*, 2015).

Applying limma is of particular advantage here, as we want to analyze both gene expression and DNA methylation data, and using the same methodological basis and statistical approach for the different types of analysis – that is differential variability in DNA methylation microarray data and differential variability in gene expression data obtained by RNAseq, both of which will be additionally compared to the results of classical analyses testing for differences in mean DNA methylation and gene expression levels – makes the results consistently comparable across all analyses performed.

A particular further strength of limma is that it takes advantage of the highly parallel nature of genomic data to borrow information between gene-wise models by the use of an empirical Bayes estimators, making the statistical conclusions more robust and reliable, especially when the number of samples is small (Ritchie *et al.*, 2015), which is typically the case in genomic analyses, where the number of tested features (that is, genes or CpGs) is orders of magnitude higher than the number of samples available.

Additionally, limma is able to incorporate a mean-variance trend into the model, which is important as both gene expression and DNA methylation values often show some degree of heteroscedasticity. The mean-variance relation is modeled via the so called "voom" ("variance modeling at the observational level") conversion (Law *et al.*, 2014). The algorithm transforms the data into normalized counts on the logarithmic scale (base 2), and estimates the relationship between mean and variance to determine precision weights that are subsequently incorporated to the linear model (Ritchie *et al.*, 2015).

Moreover, as limma is based on linear modeling, it is extremely flexible and almost any type of experimental design and customized comparisons can be handled (Ritchie *et al.*, 2015). As the data analyzed here was obtained from the same individuals for all cell types (see section 3.2.1 and figure 3.1), we performed paired tests in order to rule out confounding effects of covariates related to individuals like for example age, sex or blood cell counts. Indeed, when we added such additional information related to individuals to the paired model, only negligible changes of results could be observed. The overlap between significant results obtained by the model without additional covariates and models where distinct covariates were included was greater than 99% (data not shown).

In comparisons of commonly used methods to analyze genomic data (Seyednasrollah *et al.*, 2015; Soneson & Delorenzi, 2013; Rapaport *et al.*, 2013; Jeanmougin *et al.*, 2010; Law *et al.*, 2014), limma is often recommended as the best choice. It has proven to

perform extremely well in terms of the detection of true positives, low proportions of false positives, to deal very well with differing sequencing depths and with heterogeneous datasets, and additionally, limma is among the computationally fastest methods.

We combined the statistical framework of limma and DiffVar with further measurements of variability taking the mean-variance relationship present in DNA methylation and gene expression data into account to provide robust methodology to quantify and compare DNA methylation and gene expression variability across the three cell types investigated.

Summarizing, the integration of the well-established framework of limma with additional methods to robustly measure differential variability allows us to use the same analytical approach and statistical framework to answer different research questions and analyze distinct data types, thus making the results we obtain consistently comparable across all analysis performed in this study.

### 4.2.3 Comparison of Statistical Methods to Analyze Differential DNA Methylation Variability

First of all, as we want to analyze differences in variability between three distinct blood cell types, three group-wise comparisons leading to six lists of differentially variable loci or genes have to be made.

The comparisons:

- Monocytes versus neutrophils
- Monocytes versus T cells
- Neutrophils versus T cells

The resulting lists of significantly differentially variable sites or genes:

- Hypervariable in monocytes compared to neutrophils
- Hypervariable in monocytes compared to T cells
- Hypervariable in neutrophils compared to T cells
- Hypovariable in monocytes compared to neutrophils
- Hypovariable in monocytes compared to T cells
- Hypovariable in neutrophils compared to T cells

As stated previously, we used DiffVar (Phipson & Oshlack, 2014) for the analysis of differential variability between the three cell types. DiffVar is implemented within limma (Smyth, 2005; Ritchie *et al.*, 2015) and employs a statistical approach based on the ideas

of Levene's z-test (Levene, 1960). It calculates variability as the distance of each data point within a group from the group mean, as highly variable groups are characterized by consistently large deviations from the mean, while low variability groups show small deviations from the mean (the same concept as we used in section 4.1.6, formula 4.1). To determine if a group is significantly more variable than another, limma's t-test is then performed on the absolute deviations from the mean (MAD-values), similar to an approach presented by Jaffe *et al.* (2011), who used the MAD to assess differential variability as well.

As DiffVar is a relatively new method, we compared the results of DiffVar to those obtained by other statistical methods measuring differential variability in DNA methylation described in the literature. Bar *et al.* (2012, 2014) developed a mixture model approach to test for unequal variances between groups. This approach uses empirical Bayes modeling to borrow information across all genes similar to what is done in limma and claims to be robust to deviations from normality, which is important when dealing with DNA methylation data due to the strong bimodal distributions, as one CpG is typically either (almost) fully methylated across the population of cells measured in a methylation experiment, or mostly unmethylated (see supplementary figure SF7 in Annex I). However, the method did not identify any significant results in our data and all other DNA methylation datasets with which we have tested the method (data not shown).

Bartlett's test (Bartlett, 1937) is another statistical method to test for unequal variances that has been used for example by Teschendorff & Widschwendter (2012) to detect outliers of interest, but this test is – apart from being highly sensitive to outliers which is unwanted under many circumstances – also very sensitive to departures from normality. That is, if the data come from non-normal distributions, it may simply be testing for non-normality. The same applies for the F-test (Snedecor & Cochran, 1989), making it unsuited for testing differential variability in DNA methylation data.

As an alternative, the non-parametric Ansari-Bradley test (Ansari & Bradley, 1960) has been suggested. This procedure provides a distribution-free test of the equivalence of variances in two distributions having a common median, which is however also often not true in DNA methylation data. As stated above, methylation values are typically located close to the extremes, thus a specific site would normally be either very highly or very lowly methylated. Consequently, when there is increased variability present for a locus, which means that the values are more widely spread across the range of possible methylation values, there has to be some shift towards an intermediate mean methylation value if we want to robustly identify sites with highly variable methylation patterns compared to those with consistent methylation values across all samples of a group.

Additionally, mean methylation values lying more in the intermediate regions of the distribution of DNA methylation data also indicate heterogeneity at another level, namely at the population of the cells measured. The widely used methylation Beta-value gives the proportion of methylated probe intensities for a given CpG compared to the overall intensities measured, that is, unmethylated and methylated intensities together (see formula 3.2 in section 3.2.3), resulting in a percentage of methylation. So, if in one sample a certain CpG is methylated in exactly half of the cells measured, and unmethylated in the rest, the result would be a Beta-value of 0.5 for this locus in a given individual. As heterogeneity observed across cell populations is related to variability across individuals (see section 1.3), such intermediate methylation values do not only represent high intraindividual methylation heterogeneity but provide a further link to interindivual variability.

The Beta-value however shows heteroscedasticity, as demonstrated by Du *et al.* (2010), who found that intermediate Beta-values have increased variability compared to those close to zero or one, and as can also be seen in supplementary figure SF8. The figure shows that when methylation Beta-values are ordered according to their variance, and compared then to the corresponding mean Beta-values, different levels of variability are not evenly distributed across mean methylation values, again indicating strong heteroscedasticity, and showing that the highest variability is present for Beta-values around 0.5. As the distribution of Beta-values is bimodal (see also supplementary figure SF7), this effect can be especially well observed when splitting the data into two groups of high and low methylation Beta-values, like in supplementary figure SF9 and SF10.

While on the one hand the observation of increased variability at intermediate methylation levels makes sense, because of the previously described relationships between variability within cell populations and variability across individuals, again, the difficulty is then to determine whether increased variability is observed only because of the average methylation values lying in regions of the distribution which give higher estimates of variability per se, and could therefore mainly be caused by the quantification method used, or because the values are truly more spread out and biologically heterogeneous. Thus, measurements of variability independent of the mean are desired (the same problem as has been described in section 4.1.3 for gene expression variability analyses in CLL).

Du *et al.* (2010) suggested to use M-values instead of Beta-values in order to avoid the problem of heteroscedasticity. The M-value is the log2 ratio of the intensities of the methylated probe versus the unmethylated probe (see formula 3.3 in 3.2.3), a measure that is widely used in gene expression microarray analysis.

Indeed, when we applied for example Bartlett's test on data given in Beta-values, it identified loci as significantly differentially variable where there is actually a more dominant shift in mean methylation values present, which can be seen when looking at the corresponding M-values. One such example is demonstrated in supplementary figure SF11 in Annex I. For these reasons DiffVar only works with M-values, better suited for statistical analyses.

In our comparison of methods, presented in table 4.6, DiffVar and Bartlett's test showed quite similar numbers of significantly differentially variable sites in DNA methylation when using M-values. Around 50% of significant results were in common between these two methods (data not shown). DiffVar has been described to be generally more robust to outliers and to outperform the F-test and Bartlett's test in terms of controlling the false discovery rate (Phipson & Oshlack, 2014). Interestingly, in our analyses DiffVar identifies a higher number of significant results compared to Bartlett's test in some of the comparisons, possibly due to the increase of statistical power achieved by the enhanced empirical Bayes approach employed in limma (Ritchie *et al.*, 2015).

Table 4.6: Comparison of results obtained by different methods testing for differential variability. The numbers show the amount of statistically significant results at a FDR of 0.05. M stands for monocytes, N for neutrophils, and T for T cells. The arrow indicates hypervariability in the respective cell type of the corresponding comparison. DiffVar only works with M-values, thus no results are available for DiffVar with Beta-values.

| | | Beta-values | | | | M-values | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Bartlett | Ansari | Bar | DiffVar | Bartlett | Ansari | Bar | DiffVar |
| M vs N | N↑ | 6308 | 4 | 0 | NA | 872 | 4 | 0 | 942 |
| | M↑ | 5938 | 1 | 0 | NA | 628 | 1 | 0 | 722 |
| M vs T | T↑ | 15546 | 5 | 0 | NA | 168 | 4 | 0 | 139 |
| | M↑ | 15780 | 2 | 0 | NA | 87 | 3 | 0 | 350 |
| N vs T | T↑ | 17538 | 3 | 0 | NA | 94 | 4 | 0 | 185 |
| | N↑ | 21954 | 4 | 0 | NA | 373 | 3 | 0 | 681 |

Taking the considerations described in this section together, the framework of limma combined with DiffVar is the best choice for the complex and interconnected analyses performed in this study.

### 4.2.4 DNA Methylation Variability in Normal Blood Cells

As stated in the preceding sections 4.2.2 and 4.2.3, we used DiffVar (Phipson & Oshlack, 2014) embedded in the framework of limma (Smyth, 2005; Ritchie *et al.*, 2015) to analyze

differential variability across the three cell types. DiffVar calculates the MAD of M-values and performs the statistical test on these values. We observed that the MAD-value shows again some degree of heteroscedasticity (see figure 4.9).

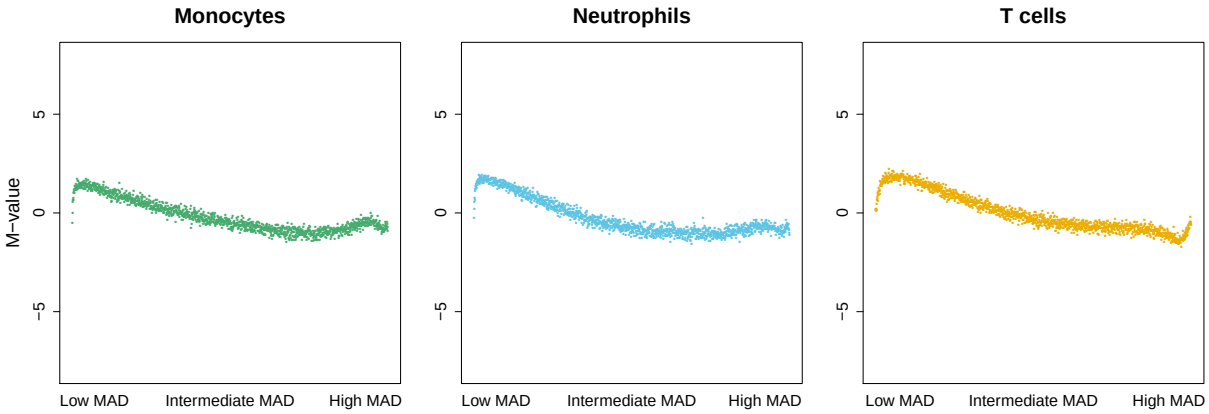


Figure 4.9: Mean M-values versus MAD. CpG-wise MAD-values were calculated. Then the values were ordered from low to high MAD, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by MAD-values, to see if the MAD is evenly distributed across M-values. Plots of the original data values in comparison with the binned data points, and the same plots separated into two groups of methylation can be seen in supplementary figures SF12, SF13 and SF14 in Annex I.

A very similar correlation between DNA methylation variability and mean methylation can be observed when looking at the variance of M-values, as shown in supplementary figures SF15, SF16 and SF17, and also reported by Heiss & Brenner (2015). Although the correlation is less pronounced here compared to the observations made on Beta-values before (see supplementary figures SF8, SF9 and SF10), these relationships between mean methylation levels and variability measurements are undesirable.

Therefore, we used an additional criterion for defining statistical significance, namely the variability measurement of Alemu *et al.* (2014) which we call MV here. The MV-score is a measurement of variability that corrects for the relationship between mean and variance, and was originally introduced to measure gene expression variability (see section 3.1.2). As we work with M-values to measure DNA methylation, similar to gene expression quantifications (Du *et al.*, 2010), and the method of Alemu *et al.* (2014) models variance as a function of the mean using the underlying data distribution, this approach can also be used here. We only considered results as significant when the Benjamini-Hochberg corrected p-values (Benjamini & Hochberg, 1995) obtained by DiffVar were smaller than 0.05, and when the absolute MV difference of a CpG between the two groups compared was bigger than 10% of the whole range of MV-values present in the methylation dataset (from -2.71 to 3.03), corresponding to an absolute MV difference of 0.57.

This way, we enforce that a significant result also has a minimum difference between the two groups in a variability measure that is less dependent on the mean. As can be seen in figure 4.10, the distribution of the MV-score indeed shows an almost flat profile, where mean methylation is centered around zero and evenly distributed across different levels of methylation variability. When again splitting up the data into two groups of high and low methylation values because of the bimodal distribution, some dependence between MV-scores and mean methylation appears, especially for extremely low DNA methylation variability (see supplementary figures SF18, SF19 and SF20), but importantly, also here the big majority of data points lies within an even region of the profile.

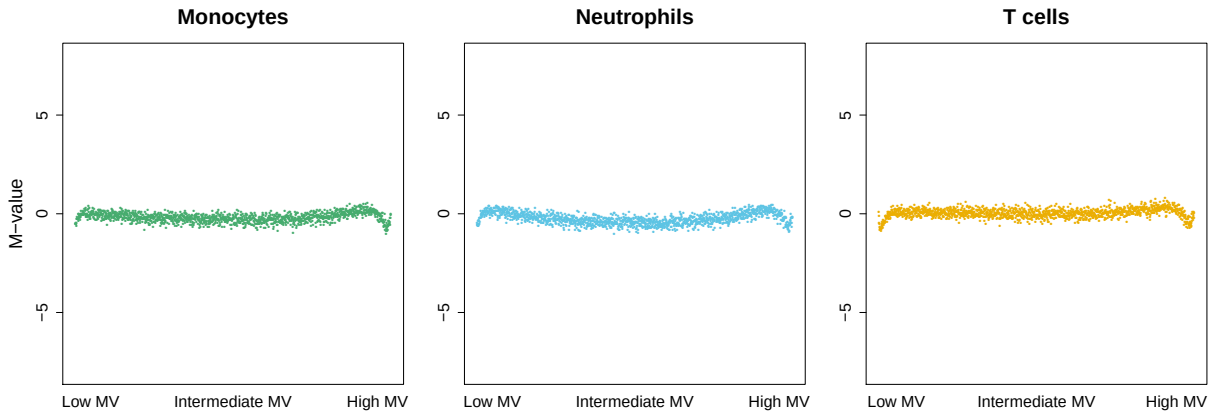


Figure 4.10: Mean M-values versus MV. CpG-wise MV-scores were calculated. Then the values were ordered from low to high MV, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by MV, to see if the MV is evenly distributed across M-values. Plots of the original data values in comparison with the binned data points, and the same plots separated into two groups of methylation can be seen in supplementary figures SF18, SF19 and SF20 in Annex I.

The numbers of significant results get only slightly reduced by applying this additional threshold in most cases – compare table 4.6 (numbers displayed in parenthesis here again) and the numbers reported in the following list – indicating that DiffVar performs generally well in identifying sites with robust differences in variability across the compared groups:

- Hypervariable in monocytes compared to neutrophils: 673 (722)
- Hypervariable in monocytes compared to T cells: 284 (350)
- Hypervariable in neutrophils compared to T cells: 449 (681)
- Hypovariable in monocytes compared to neutrophils: 713 (942)
- Hypovariable in monocytes compared to T cells: 121 (139)
- Hypovariable in neutrophils compared to T cells: 167 (185)

In order to further investigate the obtained results we defined different gene sets of interest starting from these lists of significant CpGs. First of all, we were interested in CpGs which

show cell type specific DNA methylation hypervariability, that is, CpGs presenting for example significantly increased variability in monocytes, but not in the two other cell types. To do so, we must take the two relevant comparisons into account, the comparison of monocytes versus neutrophils, and the comparison of monocytes versus T cells. If, and only if a CpG shows significantly increased variability in monocytes in both of the comparisons, it is considered as a cell type specific hypervariable locus. That means, we are looking at the overlaps of the comparisons, represented in figure 4.11.

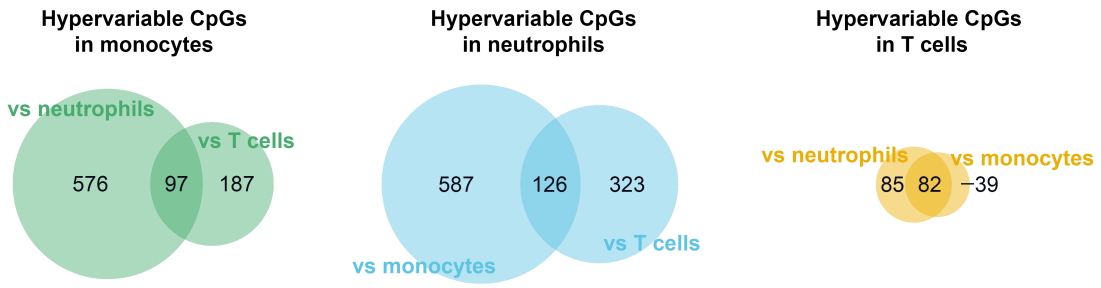


Figure 4.11: Cell type specific DNA methylation hypervariable sites. Results of the three comparisons of the analysis of differential variability represented in Venn diagrams showing the overlaps of hypervariable CpGs between two of the three comparisons in which every cell type is involved.

What can be seen here is that neutrophils show the highest number of cell type specific hypervariable sites, namely 126, as compared to 97 in monocytes, and 82 in T cells. Plotting the original methylation values of these sites together with their neighboring CpGs measured by the microarray, we can see that the procedure used to test for differential variability here indeed performs very well in identifying loci with strongly increased variability in one cell type compared to the other two. In figure 4.12 and 4.13 two interesting examples of genes with cell type specific differential methylation variability are shown.

Gene *ITGB1BP1* (integrin beta 1 binding protein 1) has two cell type specific hypervariable sites in neutrophils, a very strong one in its promoter, and another one in the 3' untranslated region (UTR). Integrins are essential cell adhesion proteins used to bind to the extracellular matrix (Alberts *et al.*, 2004; Hynes, 1987). They can regulate their affinity for extracellular ligands in order to enable cell movement, and they also function as signal transducers by the induction of intracellular signaling pathways when integrins are activated by matrix binding (Harburger & Calderwood, 2009; Miranti & Brugge, 2002). Altogether, integrins allow rapid responses to events at the cell surface (Miranti & Brugge, 2002). Specifically gene *ITGB1BP1* has also been shown to play a role in proliferation, differentiation, spreading and migration (Brunner *et al.*, 2011; Fournier *et al.*, 2005; Brütsch *et al.*, 2010).

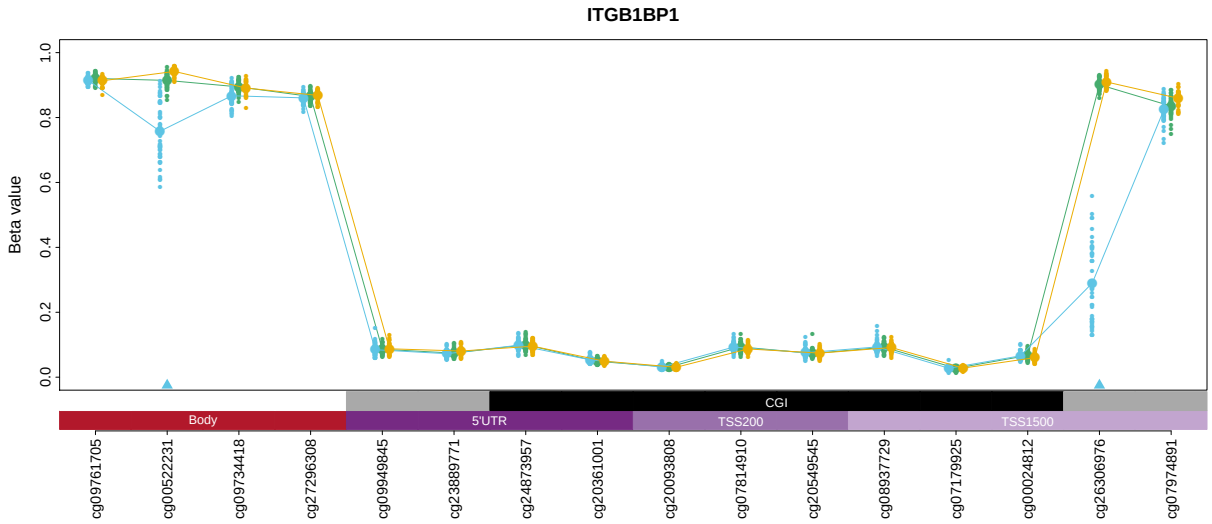


Figure 4.12: Cell type specific DNA methylation hypervariability in gene *ITGB1BP1*. Every data point represents the DNA methylation value of the CpG in one sample. The larger data points show the mean methylation value across individuals within the corresponding group, and are connected by lines. Loci with significantly increased DNA methylation variability are marked by an arrow in the color of the corresponding cell type where green stands for monocytes, blue for neutrophils, and yellow for T cells.

*HTR2A* (serotonin receptor 2A), an important neurotransmitter expressed in many cell types and playing an important role in a plethora of functions (Williams *et al.*, 2002; Van de Kar *et al.*, 2001; Tanaka *et al.*, 2008; Yu *et al.*, 2008; Hoyer *et al.*, 2002) is one of the examples showing highly variable methylation patterns across a whole region of subsequent CpGs in one of the three cell types. The promoter region of the gene shows significantly higher methylation variability in T cells than in monocytes or neutrophils.

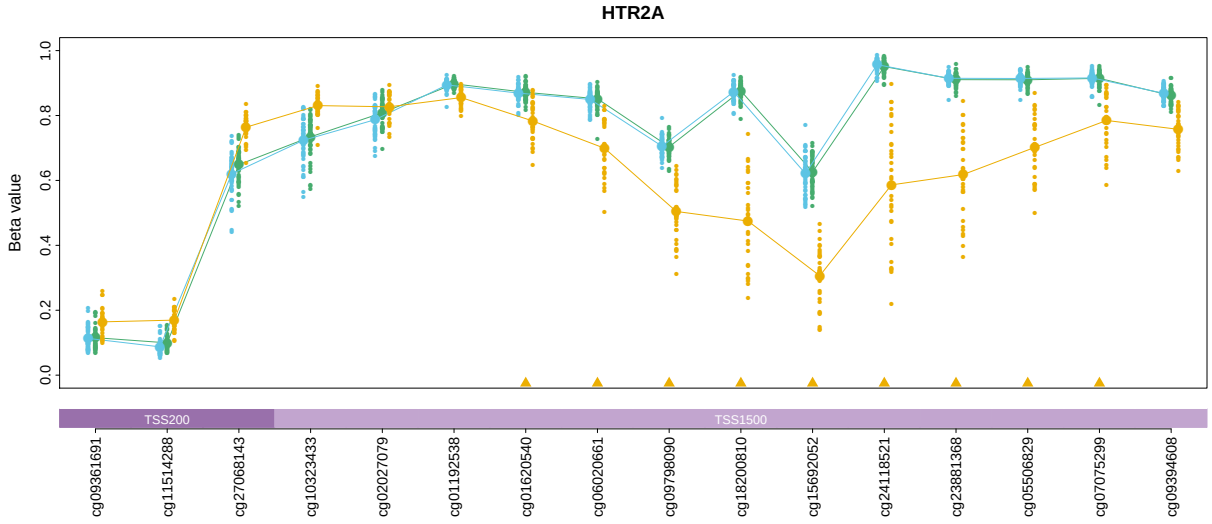


Figure 4.13: Cell type specific DNA methylation hypervariability in gene *HTR2A*. Every data point represents the DNA methylation value of the CpG in one sample. The larger data points show the mean methylation value across individuals within the corresponding group, and are connected by lines. Loci with significantly increased DNA methylation variability are marked by an arrow in the color of the corresponding cell type where green stands for monocytes, blue for neutrophils, and yellow for T cells.

Next, we wanted to see if there are hypervariable sites that are shared between two of the three cell types, which means that we want to find CpGs that have consistent methylation values in one cell type, but show highly variable methylation patterns in the other two, or said the other way round, those CpGs that are hypovariable in two of the comparisons in which the cell type in question is involved. So again, we are looking at the overlaps of the comparisons, but this time for those with decreased variability, shown in figure 4.14.

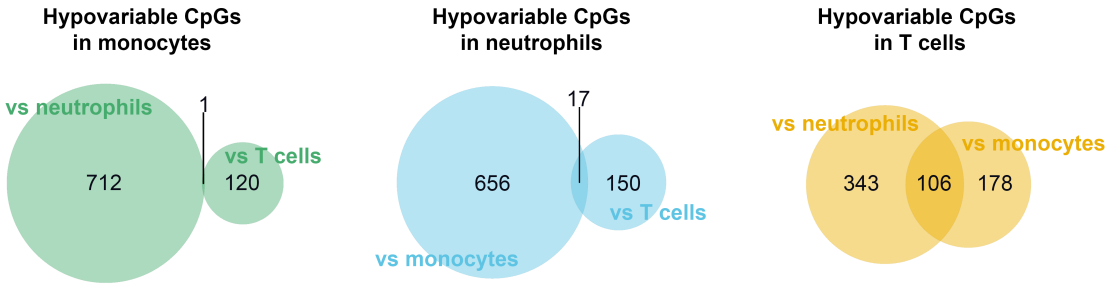


Figure 4.14: DNA methylation hypervariable sites shared between two cell types. Results of the three comparisons of the analysis of differential variability represented in Venn diagrams showing the overlaps of hypovariable CpGs between two of the three comparisons in which every cell type is involved, thus showing sites that exhibit hypervariability in the other two cell types.

The biggest overlap is present in the comparison of monocytes and neutrophils versus T cells, where 106 CpGs share hypervariability in monocytes and neutrophils and are significantly less or not variable in T cells. This is not unexpected, as monocytes and neutrophils are both cells of the myeloid lineage, and therefore more similar to each other than T cells, which belong to the lymphoid compartment.

An example gene showing hypervariable CpGs in monocytes and neutrophils but not T cells is *SOX30* (Sex Determining Region Y Box 30), shown in figure 4.15. It seem to be four subsequent probes close to the transcription start site of this gene which show such a pattern of hypervariability in both monocytes and neutrophils, but only one of them is reaching statistical significance. *SOX30* is a transcription factor involved in the regulation of embryonic development and cell fate determination (Osaki *et al.*, 1999).

Additionally, we were interested in CpGs that are highly variable in all the three cell types. Such loci cannot be identified by a statistical approach like limma, which works with contrasts between groups. Therefore, we use another method here. We took the variability measure used by DiffVar for its statistical test – the MAD-value – of all CpGs in all three cell types together, ranked the data from high variability to low variability, and took the minimum value within the top 0.05% genes as the threshold for high variability. The same was done for the MV-score. This way, we defined a MAD threshold of 1.02, and a MV threshold of 2.05.

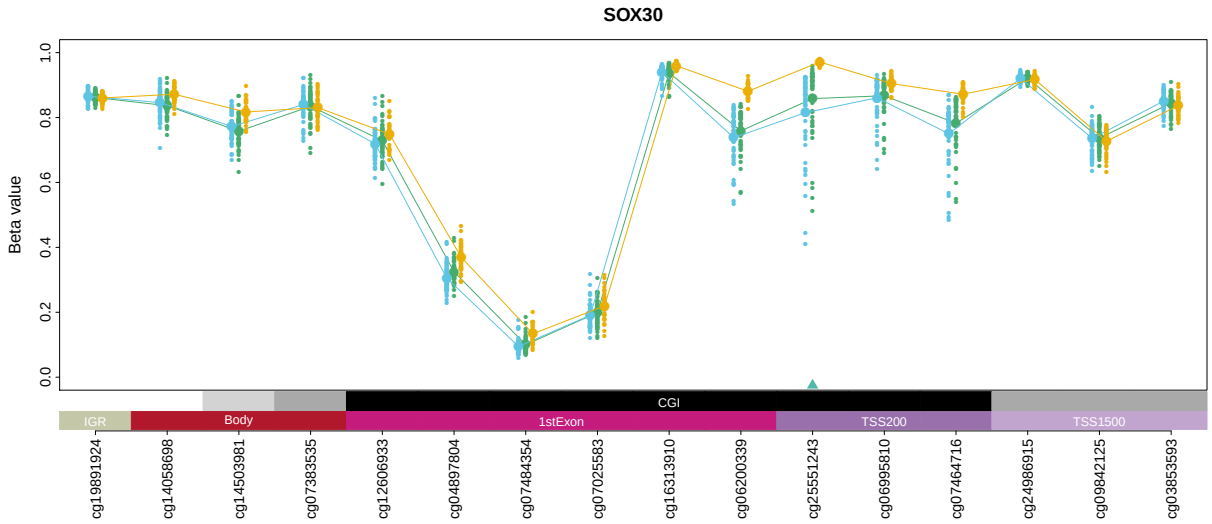


Figure 4.15: Shared DNA methylation hypervariability in gene *SOX30*. Every data point represents the DNA methylation value of the CpG in one sample. The larger data points show the mean methylation value across individuals within the corresponding group, and are connected by lines. Loci with significantly increased DNA methylation variability are marked by an arrow in the color mix of the two corresponding cell types (here, monocytes and neutrophils) where green stands for monocytes, blue for neutrophils, and yellow for T cells.

The two thresholds were then applied to all CpGs (see table 4.7), and those that showed higher variability than defined by the thresholds in all three groups were then defined as highly variable across the three cell types.

Table 4.7: Overview of CpGs with hypervariable DNA methylation patterns shared between all three cell types. The last column shows the number represented as percentage relative to all CpGs analyzed ($n = 423,089$).

| Cell type | Count | Proportion |
|---|---|---|
| Monocytes | 163 | 0.04% |
| Neutrophils | 164 | 0.04% |
| T cells | 143 | 0.03% |
| **Common to all three cell types** | **106** | **0.03%** |

Altogether, we identified 106 CpGs showing highly variable DNA methylation patterns across all three cell types. An example of such a gene, *DUSP22* (dual specificity protein phosphatase 22), is shown in figure 4.16.

*DUSP22* is an enzyme that activates the *JNK* (c-JUN N-terminal kinases) signaling pathway (Chen *et al.*, 2002; Shen *et al.*, 2001) and has been associated with several diseases such as T-cell lymphoproliferative disorders (Feldman *et al.*, 2011; Csikesz *et al.*, 2013), Alzheimer's disease (Sanchez-Mut *et al.*, 2014), and breast cancer (Sekine *et al.*, 2007), among others.

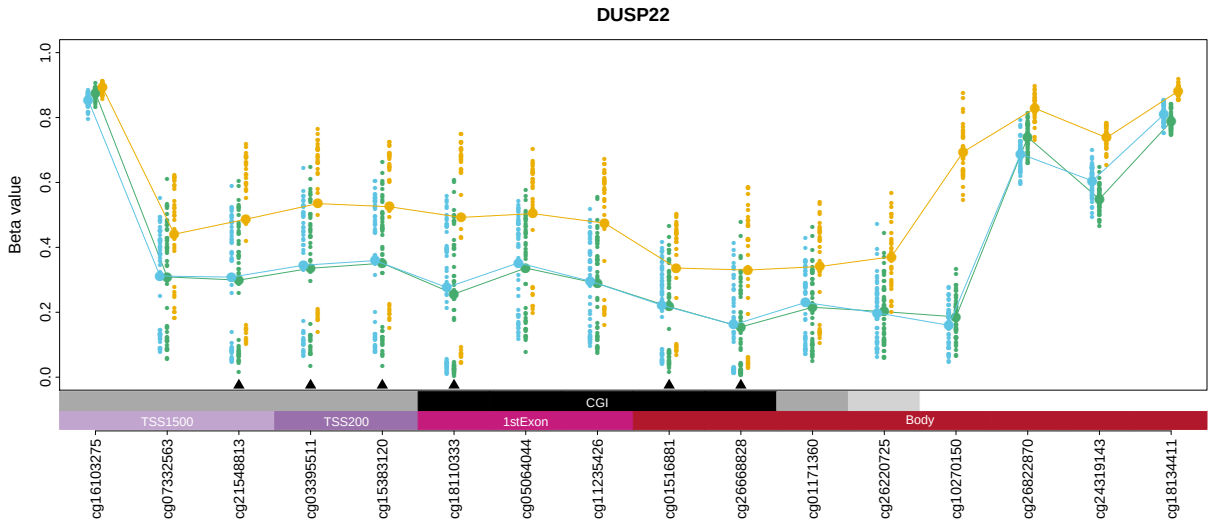


Figure 4.16: DNA methylation hypervariability common in all three cell types in gene *DUSP22*. Every data point represents the DNA methylation value of the CpG in one sample. The larger data points show the mean methylation value across individuals within the corresponding group, and are connected by lines. Loci with significantly increased DNA methylation variability in all three cell types are marked by a black arrow.

The gene shows highly variable DNA methylation values across many CpGs, in the promoter of the gene as well as in its body. However, such a pattern could also indicate technical problems with these probes on the microarray, or additional SNPs present in this region that have not been filtered out. This will have to be further investigated.

Summarizing, we found the following numbers of sites with significantly increased variability in DNA methylation, shown in figure 4.17 and listed in supplementary table ST6 table_s6.html.

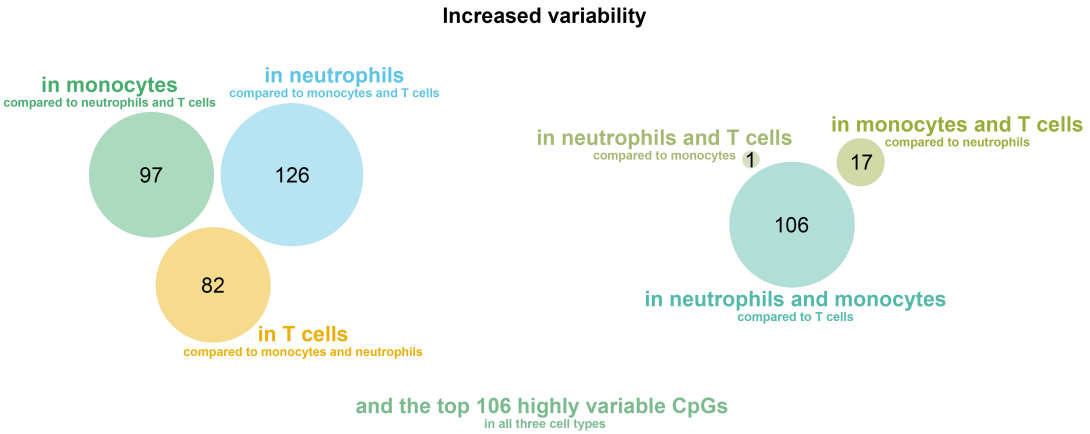


Figure 4.17: Summary of CpGs with significantly increased DNA methylation variability.

### 4.2.5 Gene Expression Variability in Normal Blood Cells

For gene expression, we performed the same analysis as described for DNA methylation variability in the preceding section 4.2.4. Also here, the MAD-value used by DiffVar is not evenly distributed across mean expression levels, and there exists a strong negative correlation between mean gene expression levels and the MAD, as shown in figure 4.18.

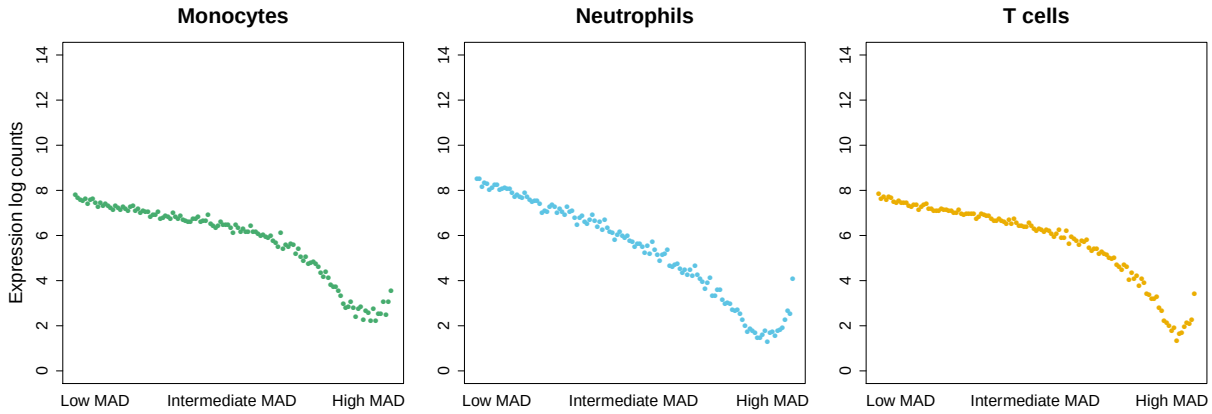


Figure 4.18: Mean expression values versus MAD. Gene-wise MAD-values were calculated. Then the values were ordered from low to high MAD, grouped together in bins of 100 genes, and plotted against mean expression log counts maintaining the ordering by MAD-values, to see if the MAD is evenly distributed across expression levels. Plots of the original data values in comparison with the binned data points can be seen in supplementary figure SF21 in Annex I.

This negative correlation between mean expression levels and variability is especially problematic here, as neutrophils exhibit generally lower expression levels than the other two cell types, which can also be seen in supplementary figure SF22 in Annex I and is consistent with the literature stating that mature neutrophils – as terminally differentiated cells – show only low transcriptional and translational activities (Geering & Simon, 2011; Amulic *et al.*, 2012; Subrahmanyam *et al.*, 2001; Wong *et al.*, 2013).

Therefore, when not correcting for this dependence, neutrophils would tend to show high variability in many genes in fact caused by overall lower expression values. We employed the previously described EV-score of Alemu *et al.* (2014) again (see section 3.1.2) and required an absolute EV difference larger than 10% of the total range of EV-values present in the expression dataset (from -2.46 to 4.41) for statistically significant results, corresponding to an absolute EV difference of at least 0.69. Furthermore, we only maintained genes for which both measurements of variability used (MAD and EV) indicated an increase of variability for the same group in the two-group comparisons. Again, the additional EV-threshold was applied to force DV genes to have a minimum difference in a measure of variability that is less dependent on mean expression levels (see figure 4.19).

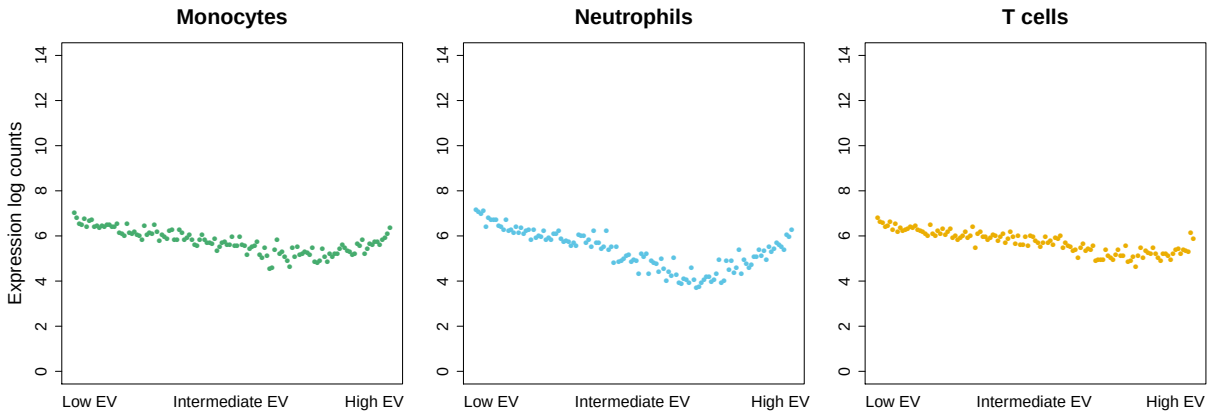


Figure 4.19: Mean expression values versus EV. Gene-wise EV-values were calculated. Then the values were ordered from low to high EV, grouped together in bins of 100 genes, and plotted against mean expression log counts maintaining the ordering by EV-values, to see if the EV is evenly distributed across expression levels. Plots of the original data values in comparison with the binned data points can be seen in supplementary figure SF23 in Annex I.

The correction for the correlation between mean expression levels and expression variability did not work perfectly here, as the EV-score still shows some dependence on the mean expression level – especially in neutrophils, where the mean expression level in the intermediate range of the EV measurement is lower – but is strongly improved compared to MAD-values, where the generally lower expression levels of neutrophils leads to a strong increase of MAD-values in this cell type (see supplementary figure SF24 in Annex I for a comparison of the global distributions of the two variability measures in the three cell types).

Indeed, when not applying the additional criterion of the minimum EV difference, more than 5,000 genes were reported to exhibit significantly increased variability in neutrophils compared to monocytes and T cells (data not shown). After the application of the additional threshold, only 544 genes reached significant levels of cell type specific hypervariability in this cell type, see figure 4.20. Still, the highest number of cell type specific hypervariability is observed for neutrophils, and there are also many genes that share their hypervariable expression patterns between monocytes and neutrophils, similar to what was observed in the analyses of differential DNA methylation variability described in section 4.2.4.

The lists of all genes with increased gene expression variability can be found in supplementary table ST7 table_s7.html.

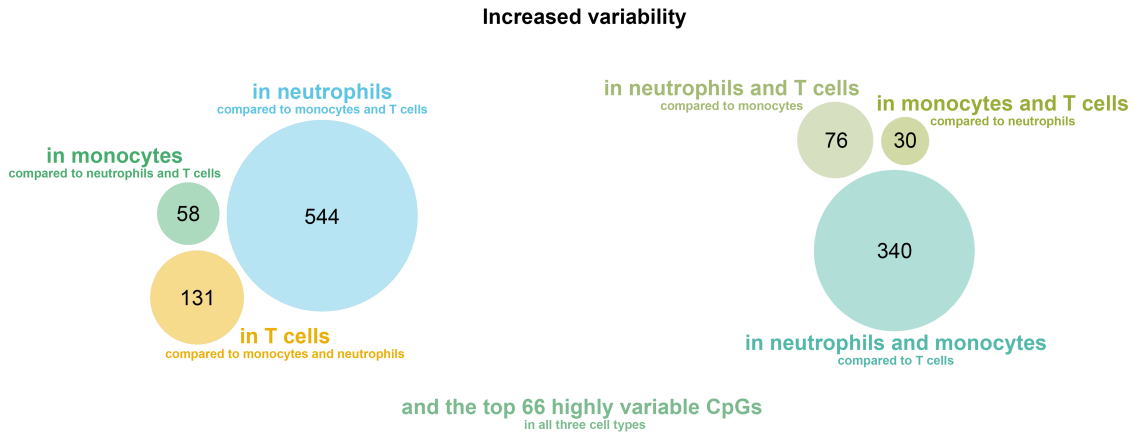


Figure 4.20: Summary of genes with significantly increased gene expression variability.

Plotting the expression values of genes with increased variability we confirmed that the differences in variability are not driven by decreased expression values in neutrophils. Some examples are shown in figure 4.21.

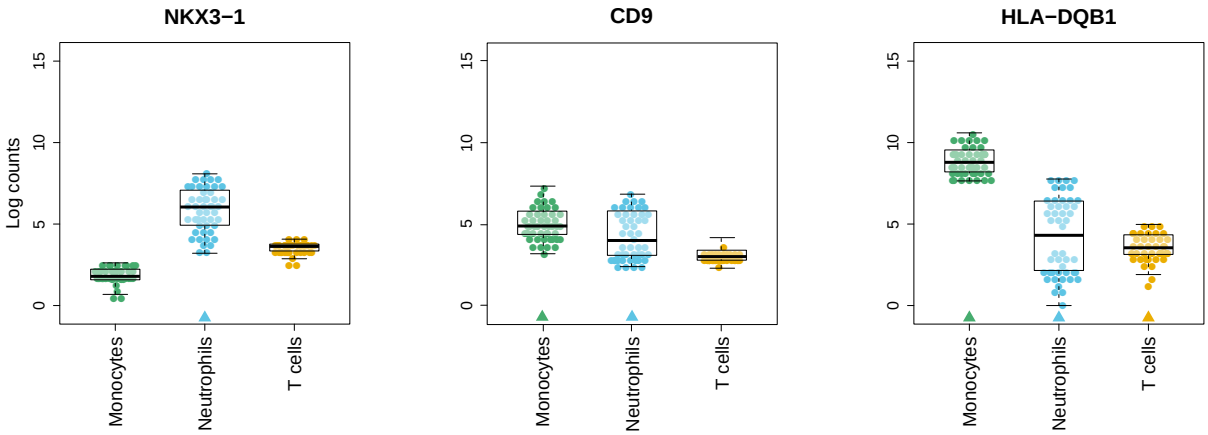


Figure 4.21: Increased gene expression variability in genes *NKX3-1*, *CD9* and *HLA-DQB1*. Every data point represents the expression value of the gene in one sample. The groups in which variability is increased are marked by an arrow in their corresponding colors.

*NKX3-1* (homeobox protein Nkx-3.1) shows increased cell type specific variability in neutrophils. The gene is an androgen-regulated transcription factor, predominantly localized to prostate epithelium with critical functions in prostate development and prostate tumor suppression (He *et al.*, 1997; Abate-Shen *et al.*, 2008; Asatiani *et al.*, 2005).

The *CD9* antigen (motility related protein 1, cell growth inhibiting gene 2 protein, or tetraspanin 29) shows hypervariable gene expression patterns in both monocytes and neutrophils, but not in T cells. It is a member of the tetraspanin family, which comprises cell surface proteins that mediate signal transduction and regulate cell development, differentiation, morphology, adhesion, activation, and motility (Hemler, 2005). *CD9* is

furthermore known to form complexes with integrins to modulate cell adhesion and migration (Hemler, 2005; Berditchevski, 2001).

Also for expression variability, we were interested in genes that share patterns of highly variable expression values across all three cell types, identified by basically the same approach as the one used for DNA methylation variability described in the previous section 4.2.4, except for the fact that here we defined the thresholds of the MAD and EV-score by looking at the top 10% of all genes, leading to a MAD cutoff of 0.45 and an EV cutoff of 0.34. The numbers of genes passing these thresholds are shown in table 4.8.

Table 4.8: Overview of genes with hypervariable expression patterns in common in all three cell types. The last column shows the number represented as percentage relative to all genes analyzed ($n = 12,661$).

| Cell type | Count | Proportion |
|---|---|---|
| Monocytes | 314 | 2.48% |
| Neutrophils | 703 | 5.55% |
| T cells | 400 | 3.16% |
| **Common to all three cell types** | **66** | **0.52%** |

An example of such a gene with hypervariable gene expression values across all three cell types is *HLA-DQB1* (major histocompatibility complex, class II, DQ beta 1), also shown in figure 4.21. This gene encodes one of the two proteins necessary to form the DQ cell surface receptor which can present antigens and bind to T cells. Therefore, the gene is primarily expressed by professional antigen-presenting cells (B cells, macrophages and dendritic cells), but can also be expressed in all other cell types (Neefjes *et al.*, 2011; Roche & Furuta, 2015).

Monocytes – as precursors of dentritic cells and macrophages – have functions associated to the presentation of antigens and possess degradative compartments enriched in major histocompatibility complex (MHC) class II molecules (Bunbury *et al.*, 2009; Hornell *et al.*, 2003). Activated neutrophils are also known to express MHC class II for antigen presentation and T cell activation (Wright *et al.*, 2010; Amulic *et al.*, 2012). Consistent with our findings here, MHC class II expression is known to be highly variable in the human population (Wright *et al.*, 2010; Gosselin *et al.*, 1993). Interestingly, the MHC class II has also been associated with the above mentioned tetraspanin proteins (Roche & Furuta, 2015) to which the *CD9* gene belongs.

In neutrophils, there seem to exist two groups of expression of *HLA-DQB1*, causing the highest intragroup variability among the three cell types. This might suggest that there

exists an important difference between individuals leading to these two different subgroups, such as for example age or sex. However, if this was the case, we would probably expect to see a similar splitting of expression values by the same individuals in monocytes and T cells as well, which does not happen. When nevertheless looking at the available phenotype data for the individuals building the two groups of expression visible in the plot, no obvious relation between the subgroups and any characteristic of the individuals for which we have information including age, sex, blood counts, date of the experiment, etc, can be found, nor can the effect be associated to any kind of further experimental variables or batch effects.

Such a grouping within cell types leading to significantly increased variability is also observable in a few other examples of statistically significant genes, and not only in neutrophils but also in the other two cell types (data not shown). It does however not seem to consistently happen with the same subgroups of individuals. Instead, the two groups of expression are formed by different individuals in different cases. This further indicates that the observation of the two groups of expression present in some genes is probably not associated with a simple phenotypic or experimental variable directly related to each individual. However, the described subgrouping pattern does only occur in few genes and is not a frequent observation among the results.

### 4.2.6 Sex-Specific Differential Expression in Normal Blood Cells

In the context of the above described observation that in some genes with hypervariable expression patterns there are two subgroups of expression levels present within a cell type (see gene *HLA-DQB1* in figure 4.21 for an example), we asked if there might exist sex-specific gene expression profiles in the three cell types investigated that could contribute to interindividual variability. Differences in mean expression levels between particular subgroups within our group of individuals such as males and females could cause significantly increased intragroup variability. We reasoned that gender could be the main phenotypic difference present.

It has to be noted that we performed the statistical test applied to assess differential variability as a paired test (see section 4.2.2), thus differences associated to individuals like age, gender or blood counts are supposed to be taken into account. However, it is important to recall that the model works on variability measurements of genes, and not the actual gene expression values, such that it accounts for effects regarding differences in variability among the individuals, but possible differences in mean expression levels corresponding to characteristics of the individuals are not (directly) taken into account.

So we asked if sex-specific differences in mean gene expression levels could be present in monocytes, neutrophils and T cells, and if yes, to which extent such differences in mean expression levels could contribute to the obtained results of highly variable gene expression patterns in the three cell types.

Using limma's student's t-test, we performed now classical differential gene expression analyses comparing mean expression levels of males and females within each of the three cell types in order to find out if there are sex-dependent differences in mean gene expression levels observable.

Surprisingly, we found 620 genes with significantly different mean expression levels between males and females in neutrophils, while only 80 and seven genes were differentially expressed between males and females in T cells and monocytes, respectively (see table 4.9 and table ST8 table_s8.html for the complete results listing all DE genes). Statistical significance was determined by Benjamini-Hochberg corrected p-values smaller than 0.05, and no fold change criterion was applied, as mean expression differences between males and females are generally small (see also supplementary table ST8 table_s8.html).

Table 4.9: Overview of genes differentially expressed between sexes.

| Cell type | Higher expressed in females | Higher expressed in males |
|---|---|---|
| Monocytes | 6 | 1 |
| Neutrophils | 299 | 321 |
| T cells | 36 | 44 |

There are no overlaps between the differentially expressed genes in the three cell types, except for one gene higher expressed in females in common between monocytes and T cells *(EIF1AXP1* – eukaryotic translation initiation factor 1A, X-linked pseudogene 1), and one gene higher expressed in males in common between monocytes and T cells *(CCDC144B* – coiled-coil domain containing 144B pseudogene).

When looking now at the overlap between the lists of genes with significantly increased gene expression variability (see section 4.2.5 and supplementary table ST7 table_s7.html) and genes exhibiting significant sex-specific differential expression, a small part of the genes indeed appears in both (see table 4.10). Thus, a fraction of the results with increased interindividual variability can at least partly be explained by sex-dependent differences in gene expression, especially in neutrophils, where the overlap between cell type specific hypervariable genes and genes with sex-dependent differences in mean expression is bigger than expected by chance according to hypergeometric tests.

Table 4.10: Overlaps of differentially variable genes with genes differentially expressed between sexes. The first column shows the number of genes with significantly differential expression between males and females in the three cell types. The second column lists the numbers of genes with increased variability, either cell type specific, shared between two cell types, or in common in all three cell types. Column 'Overlap' shows the number of genes that are contained in both lists, that is, increased gene expression variability and sex-specific differential expression in the corresponding cell type. Column 'Proportion' contains the percentage of genes with sex-specific differential expression that are present in the corresponding list of increased gene expression variability, and the last column shows the p-values of hypergeometric tests indicating if the overlaps are bigger than expected by chance.

| | | Increased variability | | Overlap | Proportion | p-value |
|---|---|---|---|---|---|---|
| Sex-specific gene expression | Neut (620) | Neutrophils | 544 | 53 | 9.74% | 0.0000 |
| | | Neutrophils & monocytes | 340 | 17 | 5.00% | 0.1384 |
| | | Neutrophils & T cells | 76 | 4 | 5.26% | 0.3168 |
| | | Common | 66 | 1 | 1.52% | 0.9196 |
| | Mono (7) | Monocytes | 58 | 0 | 0.00% | 1.0000 |
| | | Monocytes & neutrophils | 340 | 0 | 0.00% | 1.0000 |
| | | Monocytes & T cells | 30 | 0 | 0.00% | 1.0000 |
| | | Common | 66 | 0 | 0.00% | 1.0000 |
| | T cells (80) | T cells | 131 | 4 | 3.05% | 0.0037 |
| | | T cells & monocytes | 30 | 1 | 3.33% | 0.1352 |
| | | T cells & neutrophils | 76 | 0 | 0.00% | 1.0000 |
| | | Common | 66 | 0 | 0.00% | 1.0000 |

To further investigate the interesting result of 620 genes exhibiting differential expression between males and females in neutrophils, we tested for functional enrichments within the 299 genes higher expressed in females and the 321 genes higher expressed in males. Interestingly, genes that show higher expression levels in neutrophils derived from females are enriched for functions related to the immune system, while genes higher expressed in males seem to be enriched for cellular compartments like the nuclear envelope, membrane-bound organelles, and so on. The complete results of the functional enrichment analyses can be seen in supplementary table ST9 table_s9.html.

The enrichment of immune system related functions in genes that are higher expressed in females is not unexpected, as the longer life-span of females has been associated to differences in the immune system (Berghella *et al.*, 2012; Hirokawa *et al.*, 2013), it is long known that females show elevated immune responses compared to men (Grossman, 1985; Schuurs & Verheul, 1990; Ansar Ahmed *et al.*, 1985; Markle & Fish, 2014; Scotland *et al.*, 2011), and many auto-immune diseases have a higher incidence in females as well (Fairweather *et al.*, 2008; Ansar Ahmed *et al.*, 1985; Fish, 2008; Markle & Fish, 2014).

The association of genes that are higher expressed in males with cellular compartments, especially the nucleus with further enriched terms like "nuclear part" and "nuclear en-

velope" is also very interesting, as there exist differences in the polymorph nuclei of neutrophils derived from males and females (Davidson & Smith, 1954). In females, the neutrophil nucleus has an additional chromatin nodule, separated off from the main nuclear lobes. This additional structure is called neutrophil drumstick. Also sex-dependent differences in the formation of NETs have been reported (Tillack *et al.*, 2013), among other gender-specific differences in neutrophils (Aomatsu *et al.*, 2013; Molloy *et al.*, 2013; Spitzer & Zhang, 1996).

Some examples of the genes that are higher expressed in females and contribute to the enrichment of immune system processes are *CSF1* (colony stimulating factor 1), a cytokine that plays an essential role in survival, proliferation and differentiation of hematopoietic precursor cells (Stanley *et al.*, 1997), *IL-27* (Interleukin-27), a cytokine with pro- and anti-inflammatory properties that can regulate important functions in T cells (Lucas *et al.*, 2003; Pflanz *et al.*, 2002; Neufert *et al.*, 2007), and *CX3CL1* (chemokine C-X3-C motif ligand, also known as fractalkine), which has been shown to be chemotactic and plays a role in leukocyte adhesion and migration (Imai *et al.*, 1997; Bazan *et al.*, 1997).

Genes that contribute to the enrichment of the nuclear envelope cellular compartment in males are for example several nucleoporins *(NUP37, NUP155, NUP188)*, a family of proteins that build the nuclear pore complex (NPC), which is a structure that extends across the nuclear envelope and allows the flow of macromolecules between the nucleus and the cytoplasm (Doye & Hurt, 1997; Corbett & Silver, 1997), *AHCTF1* (AT hook containing transcription factor 1), a gene required for the assembly of the NPC and cell division (Rasala *et al.*, 2006; Bilokapic & Schwartz, 2012), *CSE1L* (chromosome segregation 1-like protein), an export receptor that also plays important roles in cell proliferation and apoptosis (Brinkmann *et al.*, 1995; Kutay *et al.*, 1997; Behrens *et al.*, 2003), and *LBR* (lamin B receptor), which is localized in the inner membrane of the nuclear envelope and anchors the lamina and the heterochromatin to the membrane (Pyrpasopoulou *et al.*, 1996; Ye & Worman, 1994).

For three of the genes differentially expressed between males and females in neutrophils we also found statistically significant differences in their DNA methylation patterns. The most striking example is the gene *NSD1* (nuclear receptor binding set domain protein 1), which is higher expressed in males and shows three CpGs in its promoter which are significantly hypomethylated in males as well (data not shown). Interestingly, *NSD1* has been reported to act as a transcriptional regulator and to enhance androgen receptor transactivation (Huang *et al.*, 1998). It is an autoregulatory H3 lysine-36 and H4 lyine-20 specific histone methyltransferase (Qiao *et al.*, 2011; Wang *et al.*, 2007; Lucio-Eterovic

*et al.*, 2010), and mutations of the gene cause the Sotos syndrome (Kurotaki *et al.*, 2002; Douglas *et al.*, 2003) and Weaver syndrome (Douglas *et al.*, 2003), both of which are childhood overgrowth syndromes. Furthermore, *NSD1* has been associated with acute myeloid leukemia (AML) and an adult form of the myelodysplastic syndrome (Jaju *et al.*, 2001; Wang *et al.*, 2007; Hollink *et al.*, 2011; La Starza *et al.*, 2004).

It is important to note here that the observed gender-specific differences could be confounded by hormonal alterations. In our dataset, there is a slight bias onto older individuals present in females, and most of the women included in the dataset are probably post-menopausal.

Finally, we asked if the striking differential expression between males and females in neutrophils is indeed completely neutrophil-specific. That is, if the sex-specific differential expression observed in neutrophils does only occur within neutrophils, or if it is also present in other cell types, but less strong and therefore below statistical significance. To this aim, we took the expression differences between males and females in monocytes as "baseline" in the model testing for differential expression between males and females in neutrophils. The model works by the definition of a "contrast of contrasts", comparing the differential expression observed in one contrast versus the differential expression observed in the other, with the following function call: `makeContrasts(cont=(female_neutrophils-male_neutrophils)-(female_monocytes-male_monocytes))`.

This way, the differential expression between females and males in monocytes serves as a kind of reference of expression differences to see if the differences that exist between females and males in neutrophils are significantly different from those that may be present in monocytes. If such differences between males and females are – at least partly – also present in monocytes, this approach removes such results and only reports sex-specific differential expression for genes that are truly neutrophil-specific as compared to monocytes, with the advantage of not having to rely on hard cut-offs defining statistical significance for the separate gender-comparisons of monocytes and neutrophils.

Using this technique tailored to answer the question of neutrophil-specificity, we found only 25 genes with higher expression levels in males (of which 23 are in common with the previously obtained genes) and 23 genes with higher expression in females (of which 21 are in common with previous results) in neutrophils in the current – still small – dataset, too few to perform meaningful functional enrichment analyses. Concluding, a part of the gender-specific differential expression observed in neutrophils seems to also be

present in monocytes, but less strongly pronounced, and therefore not reaching statistical significance in a classical approach testing for differential expression between males and females within each cell type separately.

### 4.2.7 Relationship Between DNA Methylation Variability and Gene Expression

Finally, we investigated the effect of DNA methylation variability on gene expression patterns using several different approaches. First of all, to get a visual impression of if there is any global tendency present, we plotted the mean expression values of each cell type versus another, and marked genes with significantly increased variability in DNA methylation. No obvious relationship between the two could be detected, genes with increased DNA methylation variability seem to be randomly distributed (see supplementary figure SF25 in Annex I).

The same happened when plotting gene expression variability represented by the EV-score. No global pattern of a relation between DNA methylation variability and expression variability could be observed (see supplementary figure SF26 in Annex I), and also not for other measurements of gene expression variability such as the CV for example (data not shown).

Next, we plotted the mean expression level and EV-score of genes with significantly differentially variable CpGs for the cell type in which the increased variability had been observed, as well as the other two cell types, in order to compare them. To this aim we used boxplots, and again, as can be seen in supplementary figure SF27 in Annex I, no consistent cell type specific or general trends are present. This might be the case because the number of genes contained in some of the boxplots is too small to produce robust patterns.

When calculating the overlaps between genes that show significantly increased DNA methylation variability in either their promoters or gene bodies and gene expression variability for the previously obtained lists of interesting genes (that is, genes with cell type specific variability, genes that share variability between two of the three cell types, and genes that show hypervariability in all three cell types), no significant overlaps could be found (see table 4.11). This result indicates that in general DNA methylation variability in a gene's promoter or body does not seem to lead to a direct and straight-forward increase of variability in its expression.

Table 4.11: Overlaps of highly variable genes in DNA methylation and gene expression. Column 'Var genes methyl' lists the number of genes with significantly increased variability in their DNA methylation, column 'Var genes exprs' lists the number of genes with significantly increased variability in their expression, column 'Overlap' represents the overlap between the two, and the last column contains the p-values of hypergeometric tests, indicating if the overlap is bigger than expected by chance.

| | | Var genes methyl | Var genes exprs | Overlap | p-value |
|---|---|---|---|---|---|
| Promoter methylation | Monocytes | 39 | 58 | 0 | 1.0000 |
| | Neutrophils | 26 | 544 | 0 | 1.0000 |
| | T cells | 18 | 131 | 0 | 1.0000 |
| | Monocytes & neutrophils | 19 | 340 | 0 | 1.0000 |
| | Monocytes & T cells | 5 | 30 | 0 | 1.0000 |
| | Neutrophils & T cells | 0 | 76 | 0 | 1.0000 |
| | In all three cell types | 19 | 66 | 0 | 1.0000 |
| Body methylation | Monocytes | 30 | 58 | 0 | 1.0000 |
| | Neutrophils | 59 | 544 | 2 | 0.8136 |
| | T cells | 31 | 131 | 2 | 0.0558 |
| | Monocytes & neutrophils | 77 | 340 | 2 | 0.7126 |
| | Monocytes & T cells | 8 | 30 | 0 | 1.0000 |
| | Neutrophils & T cells | 1 | 76 | 0 | 1.0000 |
| | In all three cell types | 61 | 66 | 1 | 0.3176 |

This observation is further confirmed by the fact that when we took the DNA methylation measurements of genes with increased interindividual methylation variability of every single individual and tried to correlate them with the corresponding expression measurements in these individuals, we often did not see any relation between DNA methylation and gene expression levels (data not shown). Similar observations have been made by Lam *et al.* (2012), who could not find an obvious association between DNA methylation variability and gene expression in their study of PBMCs across 92 individuals.

However, when analyzing the global correlation between DNA methylation variability and gene expression by plotting them against each other using a binning approach, interesting relationships between DNA methylation variability and gene expression patterns can be revealed, see figure 4.22. For the binning applied in these plots we calculated the median DNA methylation variability value per ensembl gene separately for gene promoters and gene bodies, ordered the genes by their corresponding DNA methylation variability values, binned every subsequent 100 genes together by calculating their mean expression log counts or mean expression variability, and plotted them ordered by their DNA methylation variability from low to high on the x-axis versus the corresponding variable of interest on the y-axis, either mean expression levels or expression variability measurements.

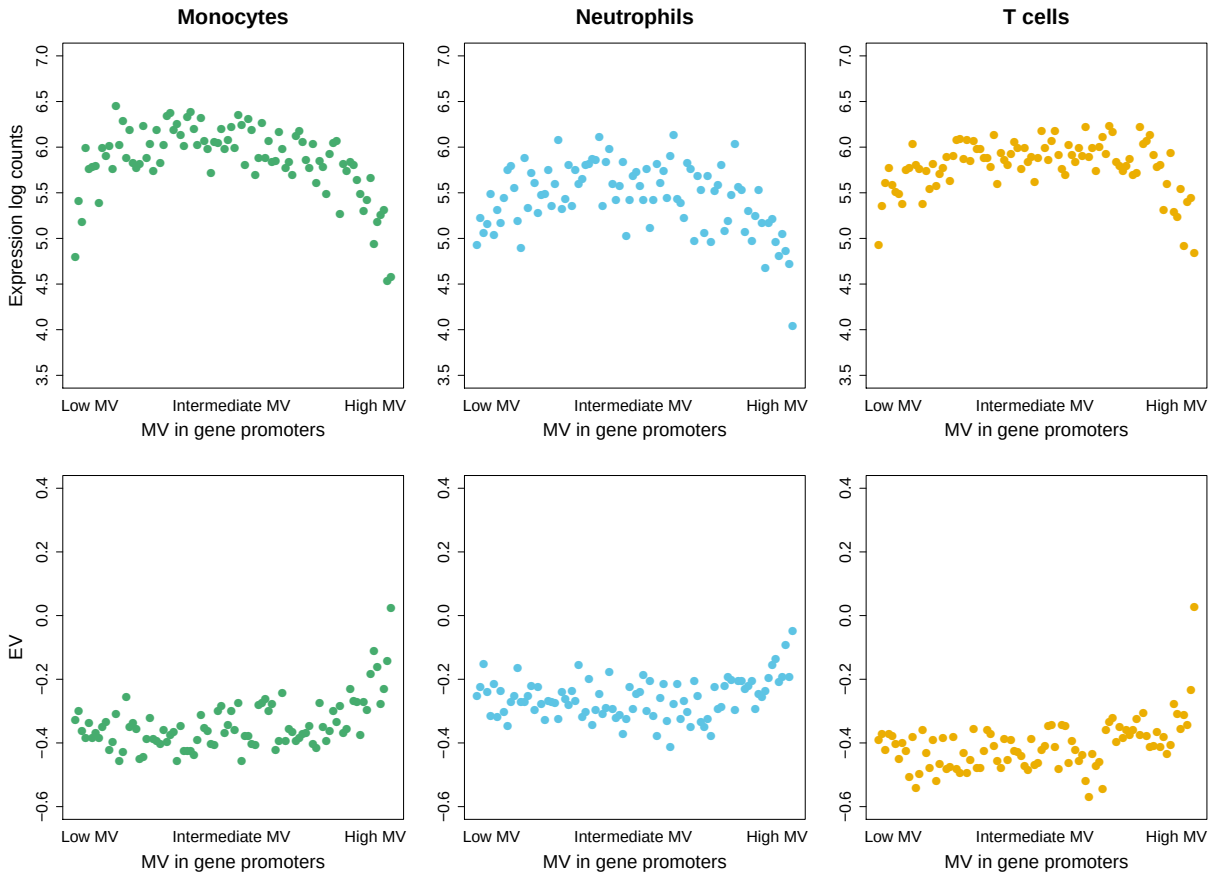


Figure 4.22: Global relationship between promoter methylation variability and gene expression. Gene-wise MV-values were calculated. Then the values were ordered from low to high MV, grouped together in bins of 100 genes, and plotted against the mean expression values (top row) or against the EV (bottom row), maintaining the ordering by MV-values. Plots of the original data values in comparison with the binned data points can be found in supplementary figure SF28 in Annex I.

This binning strategy was adopted to reduce the complexity of the data, in order to be able do detect and visualize tendencies that can be difficult to reveal when dealing with huge amounts of individual data points, as is the case here.

Indeed, when the genes were divided in classes of different levels of DNA methylation variability in their promoters as described above, a very consistent pattern emerged (see top row of figure 4.22). Genes showing very low DNA methylation variability in their promoters seem to be lowly expressed, for increasing DNA methylation variability the expression levels become higher and then lower again in a curve, and genes with extremely variable DNA methylation patterns in their promoters seem to be lowly expressed again. Strikingly, a very similar pattern can be observed for all three cell types.

When looking again at the same classes of genes – determined by the level of DNA methylation variability present in their promoters as described above – but now considering the EV-score instead of mean expression levels, a kind of inverse or mirrored pattern can be

observed, where higher expression variability is present for both very low and very high DNA methylation variability (bottom row of figure 4.22).

At first sight, the trend of an inverse pattern between mean expression levels (top row of figure 4.22) and expression variability (bottom row of figure 4.22) might lead again to the negative correlation between mean expression and expression variability. However, the EV-score is a measurement of variability that is relatively independent from mean expression levels, and the relationships that are observed when looking at figure 4.23 – which shows the same analysis for gene body methylation – do not exhibit such a mirrored image for mean expression and expression variability. This indicates that the present patterns of correlation might indeed be driven by DNA methylation variability.

For gene body methylation variability the corresponding change in mean expression or expression variability does not seem to be as strong as for promoters, and the pattern is a more linear one (see figure 4.23).

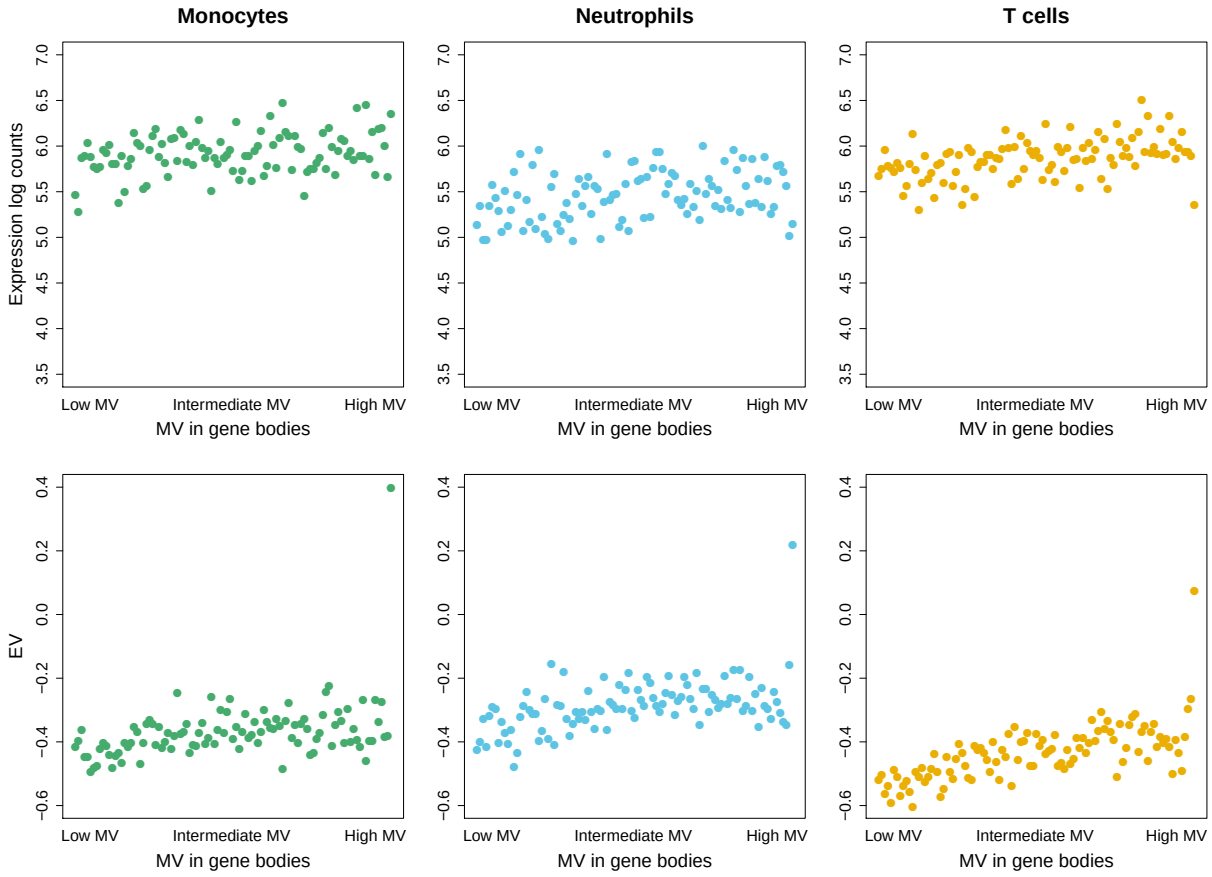


Figure 4.23: Global relationship between gene body methylation variability and gene expression. Gene-wise MV-values were calculated. Then the values were ordered from low to high MV, grouped together in bins of 100 genes, and plotted against the mean expression values (top row) or against the EV (bottom row), maintaining the ordering by MV-values. Plots of the original data values in comparison with the binned data points can be found in supplementary figure SF29 in Annex I.

Here, both increased mean expression levels and increased expression variability are observed with increasing DNA methylation variability. Especially interesting are the "outliers", at the very extremes of the plot, showing that very high DNA methylation variability in gene bodies tends to correspond to very high expression variability as well. The same seems to happen for gene promoters, where the very last data point containing the 100 genes with highest variability in promoter methylation also shows the highest level of gene expression variability (see previous figure 4.22).

### 4.2.8 Interpretation and Further Discussion

The second part of this thesis dealing with the novel field of DNA methylation and gene expression variability in normal blood cells presented here aimed to describe and discuss the methodological framework established to perform well-defined, robust and comparable analyses across different biological data types and distinct research questions.

We developed a new integrated approach combining statistical methods embedded in the well-established framework of limma (Smyth, 2005; Ritchie *et al.*, 2015) with additional measurements of variability taking the mean-variance relationship into account which is suitable for the analysis of differences in mean and differences in variability at the level of both DNA methylation and gene expression, and provides a sound basis for the study of their interrelationships.

First results showed that our approach works very well in the identification of genes with hypervariable DNA methylation and gene expression patterns, as described in section 4.2.4 and 4.2.5 and demonstrated by several examples.

Overall, our analyses on the initial dataset revealed that neutrophils exhibit both increased DNA methylation and gene expression variability compared to monocytes and T cells. This observation could possibly relate to the fact that neutrophils are the first cells to migrate to sites of inflammation (Hoffman *et al.*, 2012; Mócsai, 2013; Amulic *et al.*, 2012; Kolaczkowska & Kubes, 2013; Bardoel *et al.*, 2014), as stochastic epigenetic and transcriptional fluctuations are known to be essential to enable rapid reactions to changes in the environment, previously described in the introductory section 1.3.2 of this thesis. The observation that especially genes with important functions in intracellular signalling, cell adhesion, and motility show increased variability in neutrophils – as has also been demonstrated for some specific examples in section 4.2.4 and 4.2.5 – reinforces this hypothesis.

In recent years it has furthermore become increasingly appreciated that neutrophils are much more plastic than previously thought (Amulic *et al.*, 2012; Beyrau *et al.*, 2012; Mócsai, 2013; Takashima & Yao, 2015). It is now well-established that neutrophils play a diverse role in both innate immune defense and as effector cells of adaptive immunity, and that they use multiple highly sophisticated mechanisms to locate and kill pathogens by phagocytosis, degranulation and NETosis (Amulic *et al.*, 2012; Mócsai, 2013; Kolaczkowska & Kubes, 2013; Bardoel *et al.*, 2014). The highly variable DNA methylation and gene expression patterns of neutrophils might relate to their immense functional diversity.

Moreover, neutrophil heterogeneity from individual to individual has been previously reported for different aspects of neutrophil biology (Seligmann *et al.*, 1981; Goldschmeding *et al.*, 1992; Smith, 1994; Galli *et al.*, 2011), and there also exists the hypothesis that different functional subsets of neutrophils with distinct characteristics and biological roles might exist (Kolaczkowska & Kubes, 2013; Gallin, 1984). Differing neutrophil activation rates across individuals could also play a role in the observation of increased DNA methylation and gene expression variability in neutrophils. Additionally, it has to be noted that experimental purification procedures could have led to neutrophil activation in vitro. The gene expression patterns of neutrophils are known to change dramatically upon priming and activation (Subrahmanyam *et al.*, 2001; Wright *et al.*, 2010).

Another potential source of the observed gene expression hypervariability in neutrophils related to the specific molecular biology of this cell type might be intron retention, as it has been shown that neutrophil messenger RNAs are able to retain their introns (Wong *et al.*, 2013). The amount of retained introns increases during neutrophil differentiation, leading to strongly reduced protein levels in mature neutrophils due to intron retention dependent mechanisms of downregulation and premature stop codons (Wong *et al.*, 2013). The analyses presented in this work were performed on read counts per gene, not distinguishing between introns and exons, which could have an impact on variability observed at the level of genes. New alternative quantification techniques of RNAseq data will allow to address this hypothesis.

Beside these molecular biological differences that can potentially explain the increased variability present in neutrophils, also environmental factors could play a role here, as described in section 1.2 and 1.3.2. For example, circadian rhythms are known to have an impact on cells of the immune system (Born *et al.*, 1997; Scheiermann *et al.*, 2012; Méndez-Ferrer *et al.*, 2008; Keller *et al.*, 2009), especially also neutrophils (Wirths *et al.*, 2014; Jilma *et al.*, 1999; Abrahamsen *et al.*, 1993; Smith, 1994; Casanova-Acebes *et al.*,

2013). Additionally, seasonal changes could influence DNA methylation and gene expression variation (Dopico *et al.*, 2015; Dowell, 2001), as the samples analyzed in this work have been collected over a time span of six months.

It has further been shown that immune cells, including neutrophils, are influenced by diet, physical activity and psychological stress to name a few examples (Smith, 1994; Neubauer *et al.*, 2013; Cooper *et al.*, 2007). All these environmental factors are also known to be associated with DNA methylation and gene expression changes (Horsburgh *et al.*, 2015; Voisin *et al.*, 2015; Klengel *et al.*, 2014; Powell *et al.*, 2013; Jump & Clarke, 1999; Anderson *et al.*, 2012; Radom-Aizik *et al.*, 2008).

Summarizing, with the here developed methodology we were able to for the first time analyze interindividual DNA methylation and gene expression variability of human monocytes, neutrophils and T cells using matched samples of a sufficient number of individuals to be able to robustly quantify variability. The relation between interindividual heterogeneity and cell to cell variability introduced in the first chapter of this work allows us to link the results obtained from these analyses to the concepts of biological variability described in section 1.3.2, especially the possible relation between the observation of increased variability in neutrophils and the specific requirement of these cells to rapidly react to new or changing conditions such as the appearance of pathogens or ongoing inflammatory processes.

Notwithstanding, additional experimental validation and especially single cell sequencing data will be necessary to achieve deeper insights to verify these hypotheses and further investigate the biological implication of DNA methylation and gene expression heterogeneity in different cell types of the human immune system.

### 4.2.9 Outlook

As described in section 3.2.1, the BLUEPRINT dataset used here to develop the methodology for the diverse types of analyses performed in this project will soon be extended to 200 individuals, and will additionally also include WGS and chromatin data of the same individuals (see figure 3.1).

These data will allow us to address an important remaining aspect of this work, namely to determine to which extent the observed DNA methylation and gene expression variability can be related to genetic variation, and to separate such genetically driven variability well from heterogeneity potentially arising due to environmental factors, where especially epigenetic changes are thought to play an important role, and intrinsic stochastic variation,

as introduced in section 1.2, 1.3.1 and 1.3.2. Some authors reported that genetic heterogeneity only seemed to explain expression variability to a small extent in their studies (Li *et al.*, 2010; Battle *et al.*, 2014), and also copy number variations (CNVs) could not be directly associated to increased levels of gene expression variability (Li *et al.*, 2010; Raser & O'Shea, 2005; Alemu *et al.*, 2014).

Others however have linked gene expression variation with responsiveness to mutation (Lehner & Kaneko, 2011), and recent genome-wide quantitative trait loci mapping has revealed widespread associations of genetic variation with gene expression (Morley *et al.*, 2004; Cheung *et al.*, 2005; Göring *et al.*, 2007; Stranger *et al.*, 2007; Cheung & Spielman, 2009; Dimas *et al.*, 2009; Albert & Kruglyak, 2015; Naranbhai *et al.*, 2015; GTEx Consortium, 2015; Rivas *et al.*, 2015) and DNA methylation (Gibbs *et al.*, 2010; Bell *et al.*, 2011; Heyn *et al.*, 2013; Smith *et al.*, 2014; Wagner *et al.*, 2014; Roadmap Epigenomics Consortium, 2015).

Furthermore, the larger dataset that will soon be available from the BLUEPRINT consortium will provide the opportunity to directly address the hypotheses generated by the analyses of the pilot dataset presented in this work. The increased statistical power of the new data will for example help to further investigate the interesting result of sex-specific gene expression in neutrophils, for which only a small number of samples derived from females is available in the current dataset of 48 individuals.

The larger and more complete dataset will also allow to further interrogate the consistency of the here identified patterns of DNA methylation and gene expression variability and their interrelationships, as well as a detailed characterization thereof, for example in terms of genomic features such as CpG density, or in terms of transcription factor binding sites, which have been reported to be associated with DNA methylation variability in a study investigating mouse stem cells (Lienert *et al.*, 2011). The addition of chromatin information to the new dataset will provide important further insight here, as specific histone marks and chromatin states have been suggested to play a role in transcriptional variability as well (Busslinger & Tarakhovsky, 2014; Voss & Hager, 2014; Kaern *et al.*, 2005; Choi & Kim, 2009; Pujadas & Feinberg, 2012).

Detailed analyses of the specific biological functions of genes exhibiting hypervariable DNA methylation or gene expression patterns and functional assays to analyze for example neutrophil degranulation, respiratory burst, and adhesion in response to multiple physiological stimuli will further help to achieve a deeper understanding of epigenetic and transcriptional variability in the here investigated immune cells, and particularly in

neutrophils, for which we obtained many interesting results in the preliminary analysis of the pilot data to be followed up with the complete dataset and in further experiments.

Concluding, the present study of heterogeneity in normal human blood cells opens new doors for future research and provides grounds for additional experimental validation, which will enable a better understanding of the biological basis behind large phenotypic plasticity present in immune cells, and which will have the potential to empower the development of strategies to modulate variability in hematopoietic and immunological diseases.

# Chapter 5

# Conclusions

**1. Analysis of Variability in Chronic Lymphocytic Leukemia**

1.1 The more aggressive type of CLL, U-CLL, is characterized by higher interindividual gene expression variability.

1.2 Genes with increased gene expression variability in U-CLL are enriched for the following functions highly relevant to the disease:

(a) Intercellular communication and signaling, playing a key role in CLL through the B cell receptor.

(b) Differentiation and development, as well as cell death and apoptosis, which are basic components of leukemogenesis and disease progression.

(c) Cell cycle and proliferation, reinforcing the link between increased gene expression variability and a more progressive disease.

**2. Analysis of Variability in Normal Blood Cells**

2.1 The integrated approach combining the statistical framework limma, DiffVar, and a variability measurement that corrects for the mean-variance relationship developed here is well suited for the analysis and comparison of interindividual differential variability in DNA methylation and gene expression datasets.

2.2 Neutrophils show increased variability in their DNA methylation patterns compared to monocytes and T cells.

2.3 Neutrophils show strongly increased variability in their gene expression patterns compared to monocytes and T cells.

2.4 Neutrophils exhibit more or stronger transcriptional differences between sexes than monocytes and T cells, which partly contribute to the observed increased gene expression variability in this cell type.

2.5 There exist interesting patterns of correlation between DNA methylation variability and gene expression that are consistent across all three cell types analyzed.

**3. General Conclusion**

"Noise" matters. Epigenetic and transcriptional variability represent valuable information in the study of both physiological and pathological conditions. The here obtained results point to a crucial role of variability at the level of DNA methylation and gene expression in leukemia and the human immune system, which can be interpreted in the light of the proposed importance of intrinsic and extrinsic variability present in every biological process.

# Conclusiones

**1. Análisis de la Variabilidad en la Leucemia Linfocítica Crónica**

1.1 El tipo más agresivo de leucemia linfocítica crónica, U-CLL, se caracteriza por una mayor variabilidad de la expresión génica interindividual.

1.2 Los genes con una mayor variabilidad en la expresión génica en U-CLL están enriquecidos en las siguientes funciones de gran relevancia en la enfermedad:

(a) Comunicación y señalización intercelular, jugando un papel clave en la leucemia linfocítica crónica a través del receptor de células B.

(b) Diferenciación y desarrollo, así como muerte celular y apoptosis, las cuales son componentes básicos de la leucemogénesis y progresión de la enfermedad.

(c) Ciclo celular y proliferación, reforzando el vínculo entre el aumento de la variabilidad de la expresión génica y una enfermedad más progresiva.

**2. Análisis de la Variabilidad en Células Sanguíneas Normales**

2.1 La metodología integrada que combina el marco estadístico de limma, DiffVar, y una medida de la variabilidad que corrige por la relación entre la media y la varianza desarrollada en este trabajo, es idónea para el análisis y comparación de la variabilidad diferencial interindividual en conjuntos de datos de metilación del ADN y expresión de genes.

2.2 Los neutrófilos muestran una mayor variabilidad en sus patrones de metilación del ADN en comparación con monocitos y células T.

2.3 Los neutrófilos muestran un aumento fuerte de la variabilidad en sus patrones de expresión de genes en comparación con monocitos y células T.

2.4 Hay mayores diferencias transcripcionales entre personas de distintos sexos en neutrófilos que en monocitos y células T. Estas diferencias contribuyen parcialmente al aumento de la variabilidad de expresión génica en este tipo celular.

2.5 Existen patrones de correlación interesantes entre la variabilidad de la metilación del ADN y la expresión de genes que son consistentes en los tres tipos celulares analizados.

**3. General Conclusion**

El "ruido" es importante. La variabilidad epigenética y transcripcional representa información valiosa, tanto en el estudio de condiciones fisiológicas como patológicas. Los resultados obtenidos en este trabajo indican que la variabilidad desempeña un papel crucial a nivel de la metilación del ADN y la expresión génica en la leucemia y en el sistema inmune humano. Estos resultados pueden ser interpretados a la luz de la propuesta importancia de la variabilidad intrínseca y extrínseca presente en todos los procesos biológicos.

# List of Figures


1.1 Genes can be expressed with different efficiencies ..... 6
1.2 DNA methylation at a gene's promoter can silence its expression ..... 9
1.3 Intrinsic and extrinsic noise ..... 11
1.4 Schematic representation of the fluctuation dissipation theorem ..... 13
1.5 Hematopoiesis ..... 17
1.6 CLL cells ..... 20

3.1 Overview of the BLUEPRINT dataset of normal cells ..... 31

4.1 Differential variability versus differential mean ..... 38
4.2 Definition of the CV ..... 39
4.3 Gene expression variability comparison of U-CLL and M-CLL ..... 41
4.4 Methylation values of the top 500 genes with increased gene expression variability in U-CLL versus M-CLL ..... 43
4.5 Network of genes with increased variability in U-CLL ..... 46
4.6 Hierarchical clustering of gene expression data ..... 53
4.7 Hierarchical clustering of superpatients ..... 54
4.8 Random forest classifier results ..... 56
4.9 Mean M-values versus MAD ..... 65
4.10 Mean M-values versus MV ..... 66
4.11 Cell type specific DNA methylation hypervariable sites ..... 67
4.12 Cell type specific DNA methylation hypervariability in gene *ITGB1BP1* .. 68
4.13 Cell type specific DNA methylation hypervariability in gene *HTR2A* ..... 68
4.14 DNA methylation hypervariable sites shared between two cell types ..... 69
4.15 Shared DNA methylation hypervariability in gene *SOX30* ..... 70
4.16 DNA methylation hypervariability common in all three cell types in gene *DUSP22* ..... 71
4.17 Summary of CpGs with significantly increased DNA methylation variability 71
4.18 Mean expression values versus MAD ..... 72

4.19 Mean expression values versus EV ..... 73
4.20 Summary of genes with significantly increased gene expression variability . 74
4.21 Increased gene expression variability in genes *NKX3-1*, *CD9* and *HLA-DQB1* ..... 74
4.22 Global relationship between promoter methylation variability and gene expression ..... 83
4.23 Global relationship between gene body methylation variability and gene expression ..... 84

SF1 Scatterplots of CV and EV ..... 145
SF2 Correlation of EV and CV ..... 146
SF3 Gene expression variability comparison in additional datasets ..... 147
SF4 Correlation of variability measurements between datasets ..... 148
SF5 Correlation of EV difference and CV difference ..... 149
SF6 Beanplots comparing the methylation profiles of M-CLL and U-CLL ..... 150
SF7 Distribution of Beta-values and M-values in the three cell types ..... 153
SF8 Mean Beta-values versus variance of Beta-values ..... 154
SF9 Mean Beta-values versus variance of Beta-values in high Beta-values ..... 155
SF10 Mean Beta-values versus variance of Beta-values in low Beta-values ..... 156
SF11 Beta-values and M-values of probe cg07804973 ..... 157
SF12 Mean M-values versus MAD of M-values ..... 158
SF13 Mean M-values versus MAD of M-values in positive M-values ..... 159
SF14 Mean M-values versus MAD of M-values in negative M-values ..... 160
SF15 Mean M-values versus variance of M-values ..... 161
SF16 Mean M-values versus variance of M-values in positive M-values ..... 162
SF17 Mean M-values versus variance of M-values in negative M-values ..... 163
SF18 Mean M-values versus MV ..... 164
SF19 Mean M-values versus MV in positive M-values ..... 165
SF20 Mean M-values versus MV in negative M-values ..... 166
SF21 Mean expression values versus MAD ..... 167
SF22 Distribution of expression values in the three cell types ..... 168
SF23 Mean expression values versus EV ..... 169
SF24 Distribution of mean expression and expression variability measurements in the three cell types ..... 170
SF25 Scatterplots of mean expression levels marking genes with differential DNA methylation variability ..... 171

SF26 Scatterplots of EV-scores marking genes with differential DNA methylation variability ..... 172
SF27 Boxplots of mean expression levels and EV-scores comparing genes with cell type specific differential DNA methylation variability to others and across cell types ..... 173
SF28 Global relationship between promoter methylation variability and gene expression ..... 174
SF29 Global relationship between gene body methylation variability and gene expression ..... 175

# List of Tables


4.1 Gene expression bins ..... 39
4.2 Results of F-tests in the three datasets analyzed ..... 42
4.3 Results of hypergeometric tests assessing the overlap between differentially methylated regions and genes with increased variability in U-CLL .. 43
4.4 Functional enrichment of network modules ..... 47
4.5 Random forest classifier results ..... 57
4.6 Comparison of results obtained by different methods testing for differential variability ..... 64
4.7 Overview of CpGs with hypervariable DNA methylation patterns shared between all three cell types ..... 70
4.8 Overview of genes with hypervariable expression patterns in common in all three cell types ..... 75
4.9 Overview of genes differentially expressed between sexes ..... 77
4.10 Overlaps of differentially variable genes with genes differentially expressed between sexes ..... 78
4.11 Overlaps of highly variable genes in DNA methylation and gene expression 82

ST1 Top 500 genes with increased variablility in U-CLL ..... table_s1.html
ST2 Top 500 differentially variable genes ..... table_s2.html
ST3 Functional enrichment of hypervariable genes in U-CLL (ICGC) table_s3.html
ST4 Functional enrichment of hypervariable genes in U-CLL (Fabris) table_s4.html
ST5 Functional enrichment of network modules ..... table_s5.html
ST6 CpGs with significantly increased DNA methylation variability . table_s6.html
ST7 Genes with significantly increased gene expression variability .... table_s7.html
ST8 Sex-specific differential expression ..... table_s8.html
ST9 Functional enrichment of sex-specific differential expression ..... table_s9.html

# Bibliography

Abate-Shen, C, Shen, M & Gelmann, E. 2008. Integrating differentiation and cancer: the Nkx3.1 homeobox gene in prostate organogenesis and carcinogenesis. *Differentiation* 76(6), pp. 712–27.

Abrahamsen, JF, Smaaland, R, Sandberg, S, Aakvaag, Ar & Lote, K. 1993. Circadian variation in serum cortisol and circulating neutrophils are markers for circadian variation of bone marrow proliferation in cancer patients. *Eur J Haematol* 50(4), pp. 206–12.

Adams, D, Altucci, L, Antonarakis, SE, Ballesteros, J, Beck, S, Bird, A, Bock, C, Boehm, B, Campo, E, Caricasole, A, Dahl, F, Dermitzakis, ET, Enver, T, Esteller, M, Estivill, X, Ferguson-Smith, A, Fitzgibbon, J, Flicek, P, Giehl, C, Graf, T, Grosveld, F, Guigo, R, Gut, I, Helin, K, Jarvius, J, Küppers, R, Lehrach, H, Lengauer, T, Lernmark, A, Leslie, D, Loeffler, M, Macintyre, E, Mai, A, Martens, JH, Minucci, S, Ouwehand, WH, Pelicci, PG, Pendeville, H, Porse, B, Rakyan, V, Reik, W, Schrappe, M, Schübeler, D, Seifert, M, Siebert, R, Simmons, D, Soranzo, N, Spicuglia, S, Stratton, M, Stunnenberg, HG, Tanay, A, Torrents, D, Valencia, A, Vellenga, E, Vingron, M, Walter, J & Willcocks, S. 2012. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol* 30(3), pp. 224–6.

Albert, FW & Kruglyak, L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 16(4), pp. 197–212.

Alberts, B, Johnson, A, Lewis, J, Raff, M, Roberts, K & Walter, P. 2004. *Molecular Biology of the Cell*. New York: Garland Science, 4 ed.

Alemu, EY, Carl, JW, Corrada Bravo, H & Hannenhalli, S. 2014. Determinants of expression variability. *Nucleic Acids Res* 42(6), pp. 3503–14.

Amir, EDD, Davis, KL, Tadmor, MD, Simonds, EF, Levine, JH, Bendall, SC, Shenfeld, DK, Krishnaswamy, S, Nolan, GP & Pe'er, D. 2013. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* 31(6), pp. 545–52.

Amulic, B, Cazalet, C, Hayes, GL, Metzler, KD & Zychlinsky, A. 2012. Neutrophil Function: From Mechanisms to Disease. *Annu Rev Immunol* 30(1), pp. 459–89.

Anderson, K, Lutz, C, Van Delft, FW, Bateman, CM, Guo, Y, Colman, SM, Kempski, H, Moorman, AV, Titley, I, Swansbury, J, Kearney, L, Enver, T & Greaves, M. 2011. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* 469(7330), pp. 356–61.

Anderson, OS, Sant, KE & Dolinoy, DC. 2012. Nutrition and epigenetics: An interplay of dietary methyl donors, one-carbon metabolism and DNA methylation. *J Nutr Biochem* 23(8), pp. 853–9.

Andrews, S. 2014. *FastQC A Quality Control tool for High Throughput Sequence Data*, [Online]. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ansar Ahmed, S, Penhale, WJ & Talal, N. 1985. Sex hormones, immune responses, and autoimmune diseases. Mechanisms of sex hormone action. *Am J Pathol* 121(3), pp. 531–551.

Ansari, AR & Bradley, RA. 1960. Rank-Sum Tests for Dispersions. *Ann Mat Stat* 31(4), pp. 1174–89.

Aomatsu, M, Kato, T, Kasahara, E & Kitagawa, S. 2013. Gender difference in tumor necrosis factor-$\alpha$ production in human neutrophils stimulated by lipopolysaccharide and interferon-$\gamma$. *Biochem Biophys Res Commun* 441(1), pp. 220–5.

Aran, D, Sabato, S & Hellman, A. 2013. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol* 14(3), p. R21.

Aryee, MJ, Jaffe, AE, Corrada-Bravo, H, Ladd-Acosta, C, Feinberg, AP, Hansen, KD & Irizarry, RA. 2014. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30(10), pp. 1363–9.

Asatiani, E, Huang, WX, Wang, A, Rodriguez Ortner, E, Cavalli, LR, Haddad, BR & Gelmann, EP. 2005. Deletion, methylation, and expression of the NKX3.1 suppressor gene in primary human prostate cancer. *Cancer Res* 65(4), pp. 1164–73.

Auer, RL, Starczynski, J, McElwaine, S, Bertoni, F, Newland, AC, Fegan, CD & Cotter, FE. 2005. Identification of a potential role for POU2AF1 and BTG4 in the deletion of 11q23 in chronic lymphocytic leukemia. *Genes Chromosomes Cancer* 43(1), pp. 1–10.

Bahar, R, Hartmann, CH, Rodriguez, KA, Denny, AD, Busuttil, RA, Dollé, MET, Calder, RB, Chisholm, GB, Pollock, BH, Klein, CA & Vijg, J. 2006. Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature* 441(7096), pp. 1011–4.

Balaban, NQ, Merrin, J, Chait, R, Kowalik, L & Leibler, S. 2004. Bacterial persistence as a phenotypic switch. *Science* 305(5690), pp. 1622–5.

Bar, HY, Booth, JG & Wells, MT. 2012. A Mixture-Model Approach for Parallel Testing for Unequal Variances. *Stat Appl Genet Mol Biol* 11(1).

Bar, HY, Booth, JG & Wells, MT. 2014. A Bivariate Model for Simultaneous Testing in Bioinformatics Data. *J Am Stat Assoc* 109(506), pp. 537–47.

Bardoel, BW, Kenny, EF, Sollberger, G & Zychlinsky, A. 2014. The Balancing Act of Neutrophils. *Cell Host Microbe* 15(5), pp. 526–36.

Barkai, N & Leibler, S. 1997. Robustness in simple biochemical networks. *Nature* 387(6636), pp. 913–7.

Barski, A, Cuddapah, S, Cui, K, Roh, TY, Schones, DE, Wang, Z, Wei, G, Chepelev, I & Zhao, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* 129(4), pp. 823–37.

Bartlett, MS. 1937. Properties of Sufficiency and Statistical Tests. *Proc. R. Soc. Lond. A* 160(901), pp. 268–82.

Basehoar, AD, Zanton, SJ & Pugh, BF. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* 116(5), pp. 699–709.

Bastian, M, Heymann, S & Jacomy, M. 2009. Gephi: an open source software for exploring and manipulating networks. *Int AAAI Conf* .

Battle, A, Mostafavi, S, Zhu, X, Potash, JB, Weissman, MM, McCormick, C, Haudenschild, CD, Beckman, KB, Shi, J, Mei, R, Urban, AE, Montgomery, SB, Levinson, DF & Koller, D. 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* 24(1), pp. 14–24.

Baylin, SB & Jones, PA. 2011. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer* 11(10), pp. 726–34.

Bazan, JF, Bacon, KB, Hardiman, G, Wang, W, Soo, K, Rossi, D, Greaves, DR, Zlotnik, A & Schall, TJ. 1997. *A new class of membrane-bound chemokine with a CX3C motif.*, [Online].

Behrens, P, Brinkmann, U & Wellmann, A. 2003. CSE1L/CAS: its role in proliferation and apoptosis. *Apoptosis* 8(1), pp. 39–44.

Bell, JT, Pai, AA, Pickrell, JK, Gaffney, DJ, Pique-Regi, R, Degner, JF, Gilad, Y & Pritchard, JK. 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* 12(1), p. R10.

Benjamini, Y & Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser. B Stat Methodol* 57(1), pp. 289–300.

Berditchevski, F. 2001. Complexes of tetraspanins with integrins: more than meets the eye. *J Cell Sci* 114(Pt 23), pp. 4143–51.

Berenbaum, MC. 1972. In vivo determination of the fractional kill of human tumor cells by chemotherapeutic agents. *Cancer Chemother Rep* 56(5), pp. 563–71.

Berghella, AM, Contasta, I, Del Beato, T & Pellegrini, P. 2012. The discovery of how gender influences age immunological mechanisms in health and disease, and the identification of ageing gender-specific biomarkers, could lead to specifically tailored treatment and ultimately improve therapeutic success rates. *Immun Ageing* 9(1), p. 24.

Bergman, Y & Cedar, H. 2013. DNA methylation dynamics in health and disease. *Nat Struct Mol Biol* 20(3), pp. 274–81.

Beyrau, M, Bodkin, JV & Nourshargh, S. 2012. Neutrophil heterogeneity in health and disease: a revitalized avenue in inflammation and immunity. *Open Biol* 2(11), p. 120134.

Bilokapic, S & Schwartz, TU. 2012. Molecular basis for Nup37 and ELY5/ELYS recruitment to the nuclear pore complex. *Proc Natl Acad Sci U S A* 109(38), pp. 15241–6.

Binet, JL, Auquier, A, Dighiero, G, Chastang, C, Piguet, H, Goasguen, J, Vaugier, G, Potron, G, Colona, P, Oberling, F, Thomas, M, Tchernia, G, Jacquillat, C, Boivin, P, Lesty, C, Duault, M, Monconduit, M, Belabbes, S & Gremy, F. 1981. A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer* 48(1), pp. 198–206.

Bioconductor. 2015. *Open Source Software for Bioinformatics*, [Online]. Available at: http://www.bioconductor.org.

Bird, A. 2007. Perceptions of epigenetics. *Nature* 447(7143), pp. 396–8.

Blachly, JS, Ruppert, AS, Zhao, W, Long, S, Flynn, J, Flinn, I, Jones, J, Maddocks, K, Andritsos, L, Ghia, EM, Rassenti, LZ, Kipps, TJ, de la Chapelle, A & Byrd, JC. 2015. Immunoglobulin transcript sequence and somatic hypermutation computation from unselected RNA-seq reads in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A* 112(14), pp. 4322–7.

Blake, WJ, Balázsi, G, Kohanski, MA, Isaacs, FJ, Murphy, KF, Kuang, Y, Cantor, CR, Walt, DR & Collins, JJ. 2006. Phenotypic consequences of promoter-mediated transcriptional noise. *Mol Cell* 24(6), pp. 853–65.

Blake, WJ, KAErn, M, Cantor, CR & Collins, JJ. 2003. Noise in eukaryotic gene expression. *Nature* 422(6932), pp. 633–637.

Blausen.com. 2014. *3D rendering of various types of white blood cells*, [Online]. Available at: http://en.wikipedia.org/wiki/White_blood_cell#/media/File:Blausen_0909_WhiteBloodCells.png.

Blondel, VD, Guillaume, JL, Lambiotte, R & Lefebvre, E. 2008. Fast unfolding of communities in large networks. *J Stat Mech* P10008, pp. 1–12.

Boise, LH, González-García, M, Postema, CE, Ding, L, Lindsten, T, Turka, LA, Mao, X, Nuñez, G & Thompson, CB. 1993. Bcl-X, a Bcl-2-Related Gene That Functions As a Dominant Regulator of Apoptotic Cell Death. *Cell* 74(4), pp. 597–608.

Bonasio, R, Tu, S & Reinberg, D. 2010. Molecular signals of epigenetic states. *Science* 330(6004), pp. 612–6.

Bonavida, B, Huerta-Yepez, S, Baritaki, S, Vega, M, Liu, H, Chen, H & Berenson, J. 2011. Overexpression of Yin Yang 1 in the pathogenesis of human hematopoietic malignancies. *Crit Rev Oncog* 16(3-4), pp. 261–7.

Bonci, D, Coppola, V, Musumeci, M, Addario, A, Giuffrida, R, Memeo, L, D'Urso, L, Pagliuca, A, Biffoni, M, Labbaye, C, Bartucci, M, Muto, G, Peschle, C & De Maria, R. 2008. The miR-15a-miR-16-1 cluster controls prostate cancer by targeting multiple oncogenic activities. *Nat Med* 14(11), pp. 1271–7.

Born, J, Lange, T, Hansen, K, Mölle, M & Fehm, HL. 1997. Effects of sleep and circadian rhythm on human circulating immune cells. *J Immunol* 158(9), pp. 4454–64.

Borregaard, N. 2010. Neutrophils, from Marrow to Microbes. *Immunity* 33(5), pp. 657–70.

Bravo, HC, Pihur, V, McCall, M, Irizarry, RA & Leek, JT. 2012. Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC Bioinformatics* 13, p. 272.

Breiman, L. 2001. Random Forests. *Mach Learn* 45, pp. 5–32.

Brinkmann, U, Brinkmann, E, Gallo, M & Pastan, I. 1995. Cloning and characterization of a cellular apoptosis susceptibility gene, the human homologue to the yeast chromosome segregation gene CSE1. *Proc Natl Acad Sci U S A* 92(22), pp. 10427–31.

Brinkmann, V, Reichard, U, Goosmann, C, Fauler, B, Uhlemann, Y, Weiss, DS, Weinrauch, Y & Zychlinsky, A. 2004. Neutrophil extracellular traps kill bacteria. *Science* 303(5663), pp. 1532–5.

Brock, A, Chang, H & Huang, S. 2009. Non-genetic heterogeneity–a mutation-independent driving force for the somatic evolution of tumours. *Nat Rev Genet* 10(5), pp. 336–42.

Bruey, JM, Kantarjian, H, Ma, W, Estrov, Z, Yeh, C, Donahue, A, Sanders, H, O'Brien, S, Keating, M & Albitar, M. 2010. Circulating Ki-67 index in plasma as a biomarker and prognostic indicator in chronic lymphocytic leukemia. *Leuk Res* 34(10), pp. 1320–4.

Brunner, M, Millon-Frémillon, A, Chevalier, G, Nakchbandi, IA, Mosher, D, Block, MR, Albigès-Rizo, C & Bouvard, D. 2011. Osteoblast mineralization requires $\beta$1 integrin/ICAP-1-dependent fibronectin deposition. *J Cell Biol* 194(2), pp. 307–22.

Brütsch, R, Liebler, SS, Wüstehube, J, Bartol, A, Herberich, SE, Adam, MG, Telzerow, A, Augustin, HG & Fischer, A. 2010. Integrin cytoplasmic domain-associated protein-1 attenuates sprouting angiogenesis. *Circ Res* 107(5), pp. 592–601.

Bunbury, A, Potolicchio, I, Maitra, R & Santambrogio, L. 2009. Functional analysis of monocyte MHC class II compartments. *FASEB J* 23(1), pp. 164–71.

Busslinger, M & Tarakhovsky, A. 2014. Epigenetic control of immunity. *Cold Spring Harb Perspect Biol* 6(7), p. a024174.

Byun, HM, Siegmund, KD, Pan, F, Weisenberger, DJ, Kanel, G, Laird, PW & Yang, AS. 2009. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum Mol Genet* 18(24), pp. 4808–17.

Caligaris-Cappio, F & Hamblin, TJ. 1999. B-cell chronic lymphocytic leukemia: a bird of a different feather. *J Clin Oncol* 17(1), p. 399.

Casanova-Acebes, M, Pitaval, C, Weiss, LA, Nombela-Arrieta, C, Chèvre, R, A-González, N, Kunisaki, Y, Zhang, D, van Rooijen, N, Silberstein, LE, Weber, C, Nagasawa, T, Frenette, PS, Castrillo, A & Hidalgo, A. 2013. Rhythmic modulation of the hematopoietic niche through neutrophil clearance. *Cell* 153(5), pp. 1025–35.

Catovsky, D, Richards, S, Fooks, J & Hamblin, TJ. 1991. CLL Trials in the United Kingdom the Medical Research Council CLL Trials 1, 2 and 3. *Leuk Lymphoma* 5(S1), pp. 105–11.

Cedar, H & Bergman, Y. 2011. Epigenetics of haematopoietic cell development. *Nat Rev Immunol* 11(7), pp. 478–88.

Chan, TA, Glockner, S, Joo, MY, Chen, W, Van Neste, L, Cope, L, Herman, JG, Velculescu, V, Schuebel, KE, Ahuja, N & Baylin, SB. 2008. Convergence of mutation and epigenetic alterations identifies common genes in cancer that predict for poor prognosis. *PLoS Med* 5(5), p. e114.

Chang, HH, Hemberg, M, Barahona, M, Ingber, DE & Huang, S. 2008. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* 453(7194), pp. 544–7.

Chen, AJ, Zhou, G, Juan, T, Colicos, SM, Cannon, JP, Cabriera-Hansen, M, Meyer, CF, Jurecic, R, Copeland, NG, Gilbert, DJ, Jenkins, NA, Fletcher, F, Tan, TH & Belmont, JW. 2002. The Dual Specificity JKAP Specifically Activates the c-Jun N-terminal Kinase Pathway. *J Biol Chem* 277(39), pp. 36592–601.

Chen, N, Onisko, B & Napoli, JL. 2008. The Nuclear Transcription Factor RARalpha Associates with Neuronal RNA Granules and Suppresses Translation. *J Biol Chem* 283(30), pp. 20841–47.

Cheung, VG & Spielman, RS. 2009. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat Rev Genet* 10(9), pp. 595–604.

Cheung, VG, Spielman, RS, Ewens, KG, Weber, TM, Morley, M & Burdick, JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437(7063), pp. 1365–9.

Chiorazzi, N & Ferrarini, M. 2003. B Cell Chronic Lymphocytic Leukemia: Lessons Learned from Studies of the B Cell Antigen Receptor. *Annu Rev Immunol* 21, pp. 841–94.

Chiorazzi, N & Ferrarini, M. 2011. Cellular origin(s) of chronic lymphocytic leukemia: cautionary notes and additional considerations and possibilities. *Blood* 117(6), pp. 1781–91.

Chiorazzi, N, Rai, K & Ferrarini, M. 2005. Chronic Lymphocytic Leukemia. *N Engl J Med* 352(8), pp. 804–15.

Choi, JK & Kim, YJ. 2009. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nat Genet* 41(4), pp. 498–503.

Chong, JPJ, Thömmes, P & Blow, JJ. 1996. The role of MCM/P1 proteins in the licensing of DNA replication. *Trends Biochem Sci* 21(3), pp. 102–6.

Chuang, HY, Rassenti, L, Salcedo, M, Licon, K, Kohlmann, A, Haferlach, T, Foà, R, Ideker, T & Kipps, TJ. 2012. Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. *Blood* 120(13), pp. 2639–49.

Clark, SJ & Melki, J. 2002. DNA methylation and gene silencing in cancer: which is the guilty party? *Oncogene* 21(35), pp. 5380–7.

Claus, R, Lucas, DM, Stilgenbauer, S, Ruppert, AS, Yu, L, Zucknick, M, Mertens, D, Bühler, A, Oakes, CC, Larson, RA, Kay, NE, Jelinek, DF, Kipps, TJ, Rassenti, LZ, Gribben, JG, Dohner, H, Heerema, NA, Marcucci, G, Plass, C & Byrd, JC. 2012. Quantitative DNA methylation analysis identifies a single CpG dinucleotide important for ZAP-70 expression and predictive of prognosis in chronic lymphocytic leukemia. *J Clin Oncol* 30(20), pp. 2483–1.

CLL Trialists' Collaborative Group. 1999. Chemotherapeutic options in chronic lymphocytic leukemia: a meta-analysis of the randomized trials. CLL Trialists' Collaborative Group. *J Natl Cancer Inst* 91(10), pp. 861–8.

Cohen, AA, Geva-Zatorsky, N, Eden, E, Frenkel-Morgenstern, M, Issaeva, I, Sigal, A, Milo, R, Cohen-Saidon, C, Liron, Y, Kam, Z, Cohen, L, Danon, T, Perzov, N & Alon, U. 2008. Dynamic proteomics of individual cancer cells in response to a drug. *Science* 322(5907), pp. 1511–6.

Cooper, DM, Radom-Aizik, S, Schwindt, C & Zaldivar, F. 2007. Dangerous exercise: lessons learned from dysregulated inflammatory responses to physical activity. *J Appl Physiol* 103(2), pp. 700–9.

Corbett, AH & Silver, PA. 1997. Nucleocytoplasmic Transport of Macromolecules. *Microbiol Mol Biol Rev* 61(2), pp. 193–211.

Crick, FHC. 1970. Central Dogma of Molecular Biology. *Nature* 227, pp. 561–3.

Csikesz, CR, Knudson, RA, Greipp, PT, Feldman, AL & Kadin, M. 2013. Primary Cutaneous CD30-Positive T-Cell Lymphoproliferative Disorders with Biallelic Rearrangements of DUSP22. *J Invest Dermatol* 133(6), pp. 1680–2.

Damle, RN, Batliwalla, FM, Ghiotto, F, Valetto, A, Albesiano, E, Sison, C, Allen, SL, Kolitz, J, Vinciguerra, VP, Kudalkar, P, Wasil, T, Rai, KR, Ferrarini, M, Gregersen, PK & Chiorazzi, N. 2004. Telomere length and telomerase activity delineate distinctive replicative features of the B-CLL subgroups defined by immunoglobulin V gene mutations. *Blood* 103(2), pp. 375–82.

Damle, RN, Wasil, T, Fais, F, Ghiotto, F, Valetto, A, Allen, SL, Buchbinder, A, Budman, D, Dittmar, K, Kolitz, J, Lichtman, SM, Schulman, P, Vinciguerra, VP, Rai, KR, Ferrarini, M & Chiorazzi, N. 1999. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* 94(6), pp. 1840–7.

Davenport, J, Neale, GA & Goorha, R. 2000. Identification of genes potentially involved in LMO2-induced leukemogenesis. *Leukemia* 14(11), pp. 1986–96.

Davidson, WM & Smith, DR. 1954. A Morphological Sex Difference in the Polymorphonuclear Neutrophil Leucocytes. *Br Med J* 2(4878), pp. 6–7.

Deglesne, PA, Chevallier, N, Letestu, R, Baran-Marszak, F, Beitar, T, Salanoubat, C, Sanhes, L, Nataf, J, Roger, C, Varin-Blank, N & Ajchenbaum-Cymbalista, F. 2006. Survival response to B-cell receptor ligation is restricted to progressive chronic lymphocytic leukemia cells irrespective of Zap70 expression. *Cancer Res* 66(14), pp. 7158–66.

Dias, BG & Ressler, KJ. 2014. Parental olfactory experience influences behavior and neural structure in subsequent generations. *Nat Neurosci* 17(1), pp. 89–96.

Dimas, A, Deutsch, S, Stranger, B, Montgomery, S, Borel, C, Attar-Cohen, H, Ingle, C, Beazley, C, Gutierrez Arcelus, M, Sekowska, M, Gagnebin, M, Nisbett, J, Deloukas, P, Dermitzakis, E & Antonnarakis, S. 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325(5945), pp. 1246–50.

Dobin, A, Davis, CA, Schlesinger, F, Drenkow, J, Zaleski, C, Jha, S, Batut, P, Chaisson, M & Gingeras, TR. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1), pp. 15–21.

Döhner, H, Stilgenbauer, S, Benner, A, Leupolt, E, Kröber, A, Bullinger, L, Döhner, K, Bentz, M & Lichter, P. 2000. Genomic Aberrations and Survival in Chronic Lymphocytic Leukemia. *N Engl J Med* 343(26), pp. 1910–6.

Dominguez-Sola, D, Ying, CY, Grandori, C, Ruggiero, L, Chen, B, Li, M, Galloway, DA, Gu, W, Gautier, J & Dalla-Favera, R. 2007. Non-transcriptional control of DNA replication by c-Myc. *Nature* 448(7152), pp. 445–51.

Dong, D, Shao, X, Deng, N & Zhang, Z. 2011. Gene expression variations are predictive for stochastic noise. *Nucleic Acids Res* 39(2), pp. 403–13.

Dopico, XC, Evangelou, M, Ferreira, RC, Guo, H, Pekalski, ML, Smyth, DJ, Cooper, N, Burren, OS, Fulford, AJ, Hennig, BJ, Prentice, AM, Ziegler, AG, Bonifacio, E, Wallace, C & Todd, JA. 2015. Widespread seasonal gene expression reveals annual differences in human immunity and physiology. *Nat Commun* 6, p. 7000.

Douglas, J, Hanks, S, Temple, IK, Davies, S, Murray, A, Upadhyaya, M, Tomkins, S, Hughes, HE, Cole, TRP & Rahman, N. 2003. NSD1 mutations are the major cause of Sotos syndrome and occur in some cases of Weaver syndrome but are rare in other overgrowth phenotypes. *Am J Hum Genet* 72(1), pp. 132–3.

Dowell, SF. 2001. Seasonal Variations in Host Suceptibility and Cycles of Certain Infectious Diseases. *Emerg Infect Dis* 7(3), pp. 369–74.

Doye, V & Hurt, E. 1997. From nucleoporins to nuclear pore complexes. *Curr Opin Cell Biol* 9(3), pp. 401–11.

Du, P, Kibbe, WA & Lin, SM. 2008. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 24(13), pp. 1547–8.

Du, P, Zhang, X, Huang, CC, Jafari, N, Kibbe, WA, Hou, L & Lin, SM. 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11(1), p. 587.

Dühren-von Minden, M, Übelhart, R, Schneider, D, Wossning, T, Bach, MP, Buchner, M, Hofmann, D, Surova, E, Follo, M, Köhler, F, Wardemann, H, Zirlik, K, Veelken, H & Jumaa, H. 2012. Chronic lymphocytic leukaemia is driven by antigen-independent cell-autonomous signalling. *Nature* 489(7415), pp. 309–12.

Dvinge, H, Ries, RE, Ilagan, JO, Stirewalt, DL, Meshinchi, S & Bradley, RK. 2014. Sample processing obscures cancer-specific alterations in leukemic transcriptomes. *Proc Natl Acad Sci U S A* 111(47), pp. 16802–7.

Ecker, S. 2009. Gene Expression Analysis of T-Cell Activation. Bachelor thesis, UMIT Institute for Bioinformatics and Translational Research, Hall in Tyrol.

Ecker, S, Pancaldi, V, Rico, D & Valencia, A. 2015. Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. *Genome Med* 7(1), p. 8.

Eden, A, Gaudet, F, Waghmare, A & Jaenisch, R. 2003. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* 300(5618), p. 455.

Eferl, R & Wagner, EF. 2003. AP-1: a double-edged sword in tumorigenesis. *Nat Rev Cancer* 3(11), pp. 859–68.

Elowitz, MB, Levine, AJ, Siggia, ED & Swain, PS. 2002. Stochastic Gene Expression in a Single Cell. *Science* 297(5584), pp. 1183–6.

Engelman, JA, Zhang, X, Galbiati, F, Volonte, D, Sotgia, F, Pestell, RG, Minetti, C, Scherer, PE, Okamoto, T & Lisanti, MP. 1998. Molecular Genetics of the Caveolin Gene Family: Implications for Human Cancers, Diabetes, Alzheimer Disease, and Muscular Dystrophy. *Am J Hum Genet* 63, pp. 1578–87.

Enver, BT, Heyworth, CM & Dexter, TM. 1998. Do stem cells play dice? *Blood* 92(2), pp. 348–52.

Ernst, J & Kellis, M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 28(8), pp. 817–25.

Ernst, J, Kheradpour, P, Mikkelsen, TS, Shoresh, N, Ward, LD, Epstein, CB, Zhang, X, Wang, L, Issner, R, Coyne, M, Ku, M, Durham, T, Kellis, M & Bernstein, BE. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345), pp. 43–9.

Esteller, M. 2008. Epigenetics in Cancer. *N Engl J Med* 358(11), pp. 1148–59.

Fabris, S, Mosca, L, Todoerti, K, Cutrona, G, Lionetti, M, Intini, D, Matis, S, Colombo, M, Agnelli, L, Gentile, M, Spriano, M, Callea, V, Festini, G, Molica, S, Deliliers, GL, Morabito, F, Ferrarini, M, Neri, A & Ematologia, UO. 2008. Molecular and Transcriptional Characterization of 17p Loss in B-Cell Chronic Lymphocytic Leukemia. *Genes Chromosomes Cancer* 47(9), pp. 781–93.

Fairweather, D, Frisancho-Kiss, S & Rose, NR. 2008. Sex differences in autoimmune disease from a pathological perspective. *Am J Pathol* 173(3), pp. 600–9.

Falcon, S & Gentleman, R. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics* 23(2), pp. 257–8.

Fält, S, Merup, M, Gahrton, G, Lambert, B & Wennborg, A. 2005. Identification of progression markers in B-CLL by gene expression profiling. *Exp Hematol* 33, pp. 883–93.

Feinberg, AP & Irizarry, RA. 2010. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A* 107(Suppl 1), pp. 1757–64.

Feinberg, AP, Irizarry, RA, Fradin, D, Aryee, MJ, Gudnason, V & Fallin, MD. 2010. Personalized Epigenomic Signatures That Are Stable Over Time and Covary with Body Mass Index. *Sci Transl Med* 2(49), p. 49ra67.

Feinberg, AP & Vogelstein, B. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* 301(5895), pp. 89–92.

Feinerman, O, Veiga, J, Dorfman, JR, Germain, RN & Altan-Bonnet, G. 2008. Variability and robustness in T cell activation from regulated heterogeneity in protein levels. *Science* 321(5892), pp. 1081–4.

Feldman, AL, Dogan, A, Smith, DI, Law, ME, Ansell, SM, Johnson, SH, Porcher, JC, Özsan, N, Wieben, ED, Eckloff, BW & Vasmatzis, G. 2011. Discovery of recurrent t(6;7)(p25.3;q32.3) translocations in ALK-negative anaplastic large cell lymphomas by massively parallel genomic sequencing. *Blood* 117(3), pp. 915–20.

Ferreira, PG, Jares, P, Rico, D, Gomez-Lopez, G, Martinez-Trillos, A, Villamor, N, Ecker, S, Gonzalez-Perez, A, Knowles, DG, Monlong, J, Johnson, R, Quesada, V, Gouin, A, Djebali, S, Lopez-Guerra, M, Colomer, D, Royo, C, Cazorla, M, Pinyol, M, Clot, G, Aymerich, M, Rozman, M, Kulis, M, Tamborero, D, Papasaikas, P, Blanc, J, Gut, M, Gut, I, Puente, XS, Pisano, DG, Martin-Subero, JI, Lopez-Bigas, N, Lopez-Guillermo, A, Valencia, A, Lopez-Otin, C, Campo, E & Guigo, R. 2014. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res* 24(2), pp. 212–26.

Fish, EN. 2008. The X-files in immunity: sex-based differences predispose immune responses. *Nat Rev Immunol* 8(9), pp. 737–44.

Flynn, J, Jones, J, Johnson, AJ, Andritsos, L, Maddocks, K, Jaglowski, S, Hessler, J, Grever, MR, Ellie, I, Zhou, H, Zhu, Y, Zhang, D, Small, K, Bannerji, R & Byrd, JC. 2015. Dinaciclib is a novel cyclin dependent kinase inhibitor with significant clinical activity in relapsed and refractory chronic lymphocytic leukemia. *Leukemia* 29(7), pp. 1524–9.

Fortin, JP, Labbe, A, Lemire, M, Zanke, BW, Hudson, TJ, Fertig, EJ, Greenwood, C & Hansen, KD. 2014. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol* 15(11), p. 503.

Fournier, HN, Dupé-Manet, S, Bouvard, D, Luton, F, Degani, S, Block, MR, Retta, SF & Albiges-Rizo, C. 2005. Nuclear Translocation of Integrin Cytoplasmic Domain-associated Protein 1 Stimulates Cellular Proliferation. *Mol Biol Cell* 16(4), pp. 1859–71.

Fraga, MF, Ballestar, E, Paz, MF, Ropero, S, Setien, F, Ballestar, ML, Cigudosa, JC, Urioste, M, Benitez, J, Boix-Chornet, M, Heine-Sun, D, Sanchez-Aguilera, A, Ling, C, Carlsson, E, Poulsen, P, Vaag, A, Stephan, Z, Spector, TD, Wu, YZ, Plass, C & Esteller, M. 2005. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A* 102(30), pp. 10604–9.

Frank, DA, Mahajan, S & Ritz, J. 1997. B lymphocytes from patients with chronic lymphocytic leukemia contain signal transducer and activator of transcription (STAT) 1 and STAT3 constitutively phosphorylated on serine residues. *J Clin Invest* 100(12), pp. 3140–8.

French Cooperative Group on Chronic Lymphocytic Leukemia. 1990. Effects of chlorambucil and therapeutic decision in initial forms of chronic lymphocytic leukemia (stage A): results of a randomized clinical trial on 612 patients. *Blood* 75(7), pp. 1414–21.

Friedberg, JW, Sharman, J, Sweetenham, J, Johnston, PB, Vose, JM, LaCasce, A, Schaefer-Cutillo, J, De Vos, S, Sinha, R, Leonard, JP, Cripe, LD, Gregory, SA, Sterba, MP, Lowe, AM, Levy, R & Shipp, MA. 2010. Inhibition of Syk with fostamatinib disodium has significant clinical activity in non-Hodgkin lymphoma and chronic lymphocytic leukemia. *Blood* 115(13), pp. 2578–85.

Gahrton, G, Robèrt, KH, Friberg, K, Zech, L & Bird, AG. 1980. Extra chromosome 12 in chronic lymphocytic leukaemia. *Lancet* 315(8160), pp. 146–7.

Galli, SJ, Borregaard, N & Wynn, TA. 2011. Phenotypic and functional plasticity of cells of innate immunity: macrophages, mast cells and neutrophils. *Nat Immunol* 12(11), pp. 1035–44.

Gallin, JI. 1984. Human Neutrophil Heterogeneity Exists, But Is it Meaningful? *Blood* 63(5), pp. 977–83.

Gärtner, K. 1990. A third component causing random variability beside environment and genotype. A reason for the limited success of a 30 year long effort to standardize laboratory animals? *Lab Anim* 24(1), pp. 71–7.

Gascoigne, KE & Taylor, SS. 2008. Cancer Cells Display Profound Intra- and Interline Variation following Prolonged Exposure to Antimitotic Drugs. *Cancer Cell* 14(2), pp. 111–22.

Gautier, L, Cope, L, Bolstad, BM & Irizarry, RA. 2004. affy – analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20(3), pp. 307–15.

Geering, B & Simon, HU. 2011. Peculiarities of cell death mechanisms in neutrophils. *Cell Death Differ* 18(9), pp. 1457–69.

Geissmann, F, Manz, MG, Jung, S, Sieweke, MH, Merad, M & Ley, K. 2010. Development of Monocytes, Macrophages, and Dendritic Cells. *Science* 327(5966), pp. 656–61.

Gentleman, R. 2015. Using Categories to Model Genomic Data. Manual, Bioconductor package.

Gentleman, R, Carey, F, Huber, W, Irizarry, R & Dudoit, S. 2005. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer Science+Business Media, 1 ed.

Gerlinger, M, Rowan, AJ, Horswell, S, Larkin, J, Endesfelder, D, Gronroos, E, Martinez, P, Matthews, N, Stewart, A, Tarpey, P, Varela, I, Phillimore, B, Begum, S, Mcdonald, NQ, Butler, A, Jones, D, Raine, K, Latimer, C, Santos, CR, Nohadani, M, Eklund, AC, Spencer-Dene, B, Clark, G, Pickering, L, Stamp, G, Gore, M, Szallasi, Z, Downward, J, Futreal, PA & Swanton, C. 2012. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N Engl J Med* 366(10), pp. 883–92.

Gibbs, JR, van der Brug, MP, Hernandez, DG, Traynor, BJ, Nalls, MA, Lai, SL, Arepalli, S, Dillman, A, Rafferty, IP, Troncoso, J, Johnson, R, Zielke, HR, Ferrucci, L, Longo, DL, Cookson, MR & Singleton, AB. 2010. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* 6(5), p. e1000952.

Gilling, CE, Mittal, AK, Chaturvedi, NK, Iqbal, J, Aoun, P, Bierman, PJ, Bociek, RG, Weisenburger, DD & Joshi, SS. 2012. Lymph node-induced immune tolerance in chronic lymphocytic leukaemia: a role for caveolin-1. *Br J Haematol* 158(2), pp. 216–31.

Golding, I, Paulsson, J, Zawilski, SM & Cox, EC. 2005. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell* 123(6), pp. 1025–36.

Goldschmeding, R, Dalen, CM, Faber, N, Calafat, J, Huizinga, TWJ, Schoot, CE, Clement, LT & Borne, AEG. 1992. Further characterization of the NB 1 antigen as a variably expressed 56–62 kD GPI-linked glycoprotein of plasma membranes and specific granules of neutrophils. *Br J Haematol* 81(3), pp. 336–45.

Gordon, S & Taylor, PR. 2005. Monocyte and Macrophage Heterogeneity. *Nat Rev Immunol* 5(12), pp. 953–64.

Göring, HHH, Curran, JE, Johnson, MP, Dyer, TD, Charlesworth, J, Cole, SA, Jowett, JBM, Abraham, LJ, Rainwater, DL, Comuzzie, AG, Mahaney, MC, Almasy, L, MacCluer, JW, Kissebah, AH, Collier, GR, Moses, EK & Blangero, J. 2007. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 39(10), pp. 1208–16.

Gosselin, E, Wardwell, K, Rigby, WFC & Guyre, PM. 1993. Induction of MHC class II on human polymorphonuclear neutrophils by granulocyte/macrophage colony-stimulating factor, IFN-gamma, and IL-3. *J Immunol* 151(3), pp. 1482–90.

Greene, D, Doyle, D & Cunningham, P. 2010. Tracking the evolution of communities in dynamic social networks. In: *International conference on advances in social networks analysis and mining (ASONAM)*. IEEE, pp. 176–83.

Grossman, CJ. 1985. Interactions between the gonadal steroids and the immune system. *Science* 227(4684), pp. 257–61.

Groth, C & Lardelli, M. 2002. The structure and function of vertebrate Fibroblast Growth Factor Receptor 1. *Int J Dev Biol* 46(4), pp. 393–400.

GTEx Consortium. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348(6235), pp. 648–60.

Guantes, R, Rastrojo, A, Neves, R, Lima, A, Begoña, A & Iborra, FJ. 2015. Global variability in gene expression and alternative splicing is modulated by mitochondrial content. *Genome Res* 25(5), pp. 633–44.

Guarini, A, Chiaretti, S, Tavolaro, S, Maggio, R, Peragine, N, Citarella, F, Ricciardi, MR, Santangelo, S, Marinelli, M, Propris, MSD, Messina, M, Mauro, FR, Giudice, ID & Foa, R. 2008. BCR ligation induced by IgM stimulation results in gene expression and functional changes only in IgVH unmutated chronic lymphocytic leukemia (CLL) cells. *Blood* 112(3), pp. 782–92.

Gunnarsson, R, Mansouri, L, Isaksson, A, Göransson, H, Cahill, N, Jansson, M, Rasmussen, M, Lundin, J, Norin, S, Buhl, AM, Smedby, KE, Hjalgrim, H, Karlsson, K, Jurlander, J, Geisler, C, Juliusson, G & Rosenquist, R. 2011. Array-based genomic screening at diagnosis and during follow-up in chronic lymphocytic leukemia. *Haematologica* 96(8), pp. 1161–9.

Gurrieri, C, McGuire, P, Zan, H, Yan, XJ, Cerutti, A, Albesiano, E, Allen, SL, Vinciguerra, V, Rai, KR, Ferrarini, M, Casali, P & Chiorazzi, N. 2002. Chronic lymphocytic leukemia B cells can undergo somatic hypermutation and intraclonal immunoglobulin V(H)DJ(H) gene diversification. *J Exp Med* 196(5), pp. 629–39.

Hamblin, TJ, Davis, Z, Gardiner, A, Oscier, DG & Stevenson, FK. 1999. Unmutated IgVH Genes Are Associated With a More Aggressive Form of Chronic Lymphocytic Leukemia. *Blood* 94(6), pp. 1848–54.

Hannum, G, Guinney, J, Zhao, L, Zhang, L, Hughes, G, Sadda, S, Klotzle, B, Bibikova, M, Fan, JB, Gao, Y, Deconde, R, Chen, M, Rajapakse, I, Friend, S, Ideker, T & Zhang, K. 2013. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol Cell* 49(2), pp. 359–67.

Hansen, KD, Timp, W, Bravo, HC, Sabunciyan, S, Langmead, B, Mcdonald, OG, Wen, B, Wu, H, Liu, Y, Diep, D, Briem, E, Zhang, K, Irizarry, RA & Feinberg, AP. 2011. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 43(8), pp. 768–75.

Harburger, DS & Calderwood, DA. 2009. Integrin signalling at a glance. *J Cell Sci* 122(Pt 2), pp. 159–63.

Haslinger, C, Schweifer, N, Stilgenbauer, S, Döhner, H, Lichter, P, Kraut, N, Stratowa, C & Abseher, R. 2004. Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. *J Clin Oncol* 22(19), pp. 3937–49.

He, WW, Sciavolino, PJ, Wing, J, Augustus, M, Hudson, P, Meissner, P, Curtis, R, Shell, BK, Bostwick, DG, Tindall, DJ, Gelmann, EP, Abate-Shen, C & Carter, KC. 1997. A

Novel Human Prostate-Specific, Androgen-Regulated Homeobox Gene (NKX3.1) That Maps to 8p21, a Region Frequently Deleted in Prostate Cancer. *Genomics* 43(1), pp. 69–77.

Heiss, Ja & Brenner, H. 2015. Between-array normalization for 450K data. *Frontiers in Genetics* 6(March), pp. 1–7.

Helin, K & Dhanak, D. 2013. Chromatin proteins and modifications as drug targets. *Nature* 502(7472), pp. 480–8.

Hemler, ME. 2005. Tetraspanin functions and associated microdomains. *Nat Rev Mol Cell Biol* 6(10), pp. 801–11.

Herishanu, Y, Pérez-Galán, P, Liu, D, Biancotto, A, Pittaluga, S, Vire, B, Gibellini, F, Njuguna, N, Lee, E, Stennett, L, Raghavachari, N, Liu, P, McCoy, JP, Raffeld, M, Stetler-Stevenson, M, Yuan, C, Sherry, R, Arthur, DC, Maric, I, White, T, Marti, GE, Munson, P, Wilson, WH & Wiestner, A. 2011. The lymph node microenvironment promotes B-cell receptor signaling, NF-kappaB activation, and tumor proliferation in chronic lymphocytic leukemia. *Blood* 117(2), pp. 563–74.

Heyn, H, Li, N, Ferreira, HJ, Moran, S, Pisano, DG, Gomez, A, Diez, J, Sanchez-Mut, JV, Setien, F, Carmona, FJ, Puca, AA, Sayols, S, Pujana, MA, Serra-Musach, J, Iglesias-Platas, I, Formiga, F, Fernandez, AF, Fraga, MF, Heath, SC, Valencia, A, Gut, IG, Wang, J & Esteller, M. 2012. Distinct DNA methylomes of newborns and centenarians. *Proc Natl Acad Sci U S A* 109(26), pp. 10522–7.

Heyn, H, Moran, S, Hernando-Herraez, I, Sayols, S, Gomez, A, Sandoval, J, Monk, D, Hata, K, Marques-bonet, T, Wang, L & Esteller, M. 2013. DNA methylation contributes to natural human variation DNA methylation contributes to natural human variation. *Genome Res* 23(9), pp. 1363–72.

Hirokawa, K, Utsuyama, M, Hayashi, Y, Kitagawa, M, Makinodan, T & Fulop, T. 2013. Slower immune system aging in women versus men in the Japanese population. *Immun Ageing* 10(1), p. 19.

Ho, JWK, Stefani, M, Dos Remedios, CG & Charleston, MA. 2008. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics* 24(13), pp. i390–8.

Hoffbrand, AV, Pettit, JE & Moss, PAH. 2005. *Essential Haematology*. London: Blackwell Science, 4 ed.

Hoffman, R, Benz Jr, EJ, Silberstein, LE, Heslop, H, Weitz, J & Anastasi, J. 2012. *Neutrophil structure and function. Hematology: Basic Principles and Practice*. Elsevier Health Sciences.

Holliday, R. 1987. The Inheritance of Epigenetic Defects. *Science* 238(11), pp. 163–70.

Holliday, R & Pugh, JE. 1975. DNA modification mechanisms and gene activity during development. *Science* 187(4173), pp. 226–32.

Hollink, IHIM, Heuvel-Eibrink, MMVD, Arentsen-Peters, STCJM, Pratcorona, M, Abbas, S, Kuipers, JE, Galen, JFV, Beverloo, HB, Sonneveld, E, Kaspers, GJJL, Trka, J, Baruchel, A, Zimmermann, M, Creutzig, U, Reinhardt, D, Pieters, R, Valk, PJM & Zwaan, CM. 2011. NUP98/NSD1 characterizes a novel poor prognostic group in acute myeloid leukemia with a distinct HOX gene expression pattern. *Blood* 118(13), pp. 3645–56.

Hornell, TMC, Beresford, GW, Bushey, A, Boss, JM & Mellins, ED. 2003. Regulation of the class II MHC pathway in primary human monocytes by granulocyte-macrophage colony-stimulating factor. *J Immunol* 171(5), pp. 2374–83.

Horsburgh, S, Robson-Ansley, P, Adams, R & Smith, C. 2015. Exercise and inflammation-related epigenetic modifications: focus on DNA methylation. *Exerc Immunol Rev* 21, pp. 26–41.

Horvath, S. 2013. DNA methylation age of human tissues and cell types. *Genome Biol* 14(10), p. R115.

Hovestadt, V, Jones, DTW, Picelli, S, Wang, W, Kool, M, Northcott, PA, Sultan, M, Stachurski, K, Ryzhova, M, Warnatz, HJ, Ralser, M, Brun, S, Bunt, J, Jäger, N, Kleinheinz, K, Erkek, S, Weber, UD, Bartholomae, CC, von Kalle, C, Lawerenz, C, Eils, J, Koster, J, Versteeg, R, Milde, T, Witt, O, Schmidt, S, Wolf, S, Pietsch, T, Rutkowski, S, Scheurlen, W, Taylor, MD, Brors, B, Felsberg, J, Reifenberger, G, Borkhardt, A, Lehrach, H, Wechsler-Reya, RJ, Eils, R, Yaspo, ML, Landgraf, P, Korshunov, A, Zapatka, M, Radlwimmer, B, Pfister, SM & Lichter, P. 2014. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* 510(7506), pp. 537–41.

Hoxha, M, Fabris, S, Agnelli, L, Bollati, V, Cutrona, G, Matis, S, Recchia, AG, Gentile, M, Cortelezzi, A, Morabito, F, Bertazzi, PA, Ferrarini, M & Neri, A. 2014. Relevance of telomere/telomerase system impairment in early stage chronic lymphocytic leukemia. *Genes Chromosomes Cancer* 53(7), pp. 612–21.

Hoyer, D, Hannon, JP & Martin, GR. 2002. Molecular, pharmacological and functional diversity of 5-HT receptors. *Pharmacol Biochem Behav* 71(4), pp. 533–54.

Huang, DW, Sherman, BT & Lempicki, RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1), pp. 44–57.

Huang, N, Vom Baur, E, Garnier, JM, Lerouge, T, Vonesch, JL, Lutz, Y, Chambon, P & Losson, R. 1998. Two distinct nuclear receptor interaction domains in NSD1, a novel SET protein that exhibits characteristics of both corepressors and coactivators. *EMBO J* 17(12), pp. 3398–412.

Hulse, AM & Cai, JJ. 2013. Genetic variants contribute to gene expression variability in humans. *Genetics* 193(1), pp. 95–108.

Hume, DA. 2000. Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. *Blood* 96(7), pp. 2323–8.

Hynes, R. 1987. Integrins: a family of cell adhesion receptors. *Cell* 48(4), pp. 549–54.

Iacovelli, S, Hug, E, Bennardo, S, Duehren-von Minden, M, Gobessi, S, Rinaldi, A, Suljagic, M, Bilbao, D, Bolasco, G, Eckl-Dorna, J, Niederberger, V, Autore, F, Sica, S, Laurenti, L, Wang, H, Cornall, R, Clarke, S, Croce, C, Bertoni, F, Jumaa, H & Efremov, D. 2015. Two types of BCR interactions are positively selected during leukemia development in the E$\mu$-TCL1 transgenic mouse model of CLL. *Blood* 125(10), pp. 1578–88.

Illumina Inc. 2015. *HumanMethylation450 v1.2 Manifest File*, [Online]. Available at: http://support.illumina.com/downloads/infinium_humanmethylation450_product_files.html. Accessed 2015-01-29.

Imai, T, Hieshima, K, Haskell, C, Baba, M, Nagira, M, Nishimura, M, Kakizaki, M, Takagi, S, Nomiyama, H, Schall, TJ & Yoshie, O. 1997. Identification and molecular characterization of fractalkine receptor CX3CR1, which mediates both leukocyte migration and adhesion. *Cell* 91(4), pp. 521–30.

International Cancer Genome Consortium. 2010. International network of cancer genome projects. *Nature* 464(7291), pp. 993–8.

International Cancer Genome Consortium. 2015. *ICGC Cancer Genome Project - Spain - Chronic Lymphocytic Leukemia*, [Online]. Available at: https://icgc.org/icgc/cgp/64/530/826.

Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U & Speed, TP. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2), pp. 249–64.

Issa, JP. 2011. Epigenetic variation and cellular Darwinism. *Nat Genet* 43(8), pp. 724–6.

Jaffe, AE, Feinberg, AP, Irizarry, RA & Leek, JT. 2011. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics* 13(1), pp. 166–78.

Jaju, RJ, Fidler, C, Haas, OA, Strickson, AJ, Watkins, F, Clark, K, Cross, NCP, Cheng, JF, Aplan, PD, Kearney, L, Boultwood, J & Wainscoat, JS. 2001. A novel gene, NSD1, is fused to NUP98 in the t(5;11)(q35;p15.5) in de novo childhood acute myeloid leukemia. *Blood* 98(4), pp. 1264–7.

Janeway, CA, Travers, P & Walport, M. 2001. *Immunobiology: the immune system in health and disease*. New York: Garland Science, 5 ed.

Jeanmougin, M, de Reynies, A, Marisa, L, Paccard, C, Nuel, G & Guedj, M. 2010. Should we abandon the t-Test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS One* 5(9), p. e12336.

Jenuwein, T & Allis, C. 2001. Translating the Histone Code. *Science* 293(5532), pp. 1074–80.

Jiang, A & Clark, EA. 2001. Involvement of Bik, a Proapoptotic Member of the Bcl-2 Family, in Surface IgM-Mediated B Cell Apoptosis. *J Immunol* 166(10), pp. 6025–33.

Jilma, B, Hergovich, N, Stohlawetz, P, Eichler, H, Bauer, P & Wagner, O. 1999. Circadian variation of granulocyte colony-stimulating factor levels in man. *Br J Haematol* 106(2), pp. 368–70.

Johnson, WE, Li, C & Rabinovic, A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1), pp. 118–27.

Jones, PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 13(7), pp. 484–92.

Jones, PA & Baylin, SB. 2002. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 3(6), pp. 415–28.

Juliusson, G & Gahrton, G. 1993. Cytogenetics in CLL and related disorders. *Baillieres Clin Haematol* 6(4), pp. 821–48.

Jump, DB & Clarke, SD. 1999. Regulation of gene expression by dietary fat. *Annu Rev Nutr* 19(1), pp. 63–90.

Kaern, M, Elston, TC, Blake, WJ & Collins, JJ. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* 6(6), pp. 451–64.

Kampen, NGV. 2007. *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library.

Kanduri, M, Cahill, N, Göransson, H, Enström, C, Ryan, F, Isaksson, A & Rosenquist, R. 2010. Differential genome-wide array - Based methylation profiles in prognostic subsets of chronic lymphocytic leukemia. *Blood* 115(2), pp. 296–305.

Keller, M, Mazuch, J, Abraham, U, Eom, GD, Herzog, ED, Volk, HD, Kramer, A & Maier, B. 2009. A circadian clock in macrophages controls inflammatory immune responses. *Proc Natl Acad Sci U S A* 106(50), pp. 21407–12.

Kellogg, RA & Tay, S. 2015. Noise Facilitates Transcriptional Control under Dynamic Inputs. *Cell* 160(3), pp. 381–92.

Kierzek, AM, Zaim, J & Zielenkiewicz, P. 2001. The Effect of Transcription and Translation Initiation Frequencies on the Stochastic Fluctuations in Prokaryotic Gene Expression. *J Biol Chem* 276(11), pp. 8165–72.

Kipps, TJ. 2007. The B-cell receptor and ZAP-70 in chronic lymphocytic leukemia. *Best Pract Res Clin Haematol* 20(3), pp. 415–424.

Klein, U & Dalla-Favera, R. 2008. Germinal centres: role in B-cell physiology and malignancy. *Nat Rev Immunol* 8(1), pp. 22–33.

Klein, U, Tu, Y, Stolovitzky, GA, Mattioli, M, Cattoretti, G, Husson, H, Freedman, A, Inghirami, G, Cro, L, Baldini, L, Neri, A, Califano, A & Dalla-Favera, R. 2001. Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J Exp Med* 194(11), pp. 1625–38.

Klengel, T, Pape, J, Binder, EB & Mehta, D. 2014. The role of DNA methylation in stress-related psychiatric disorders. *Neuropharmacology* 80, pp. 115–32.

Kleppe, M & Levine, RL. 2014. Tumor Heterogeneity Confounds and Illuminates: Assessing the implications. *Nat Med* 20(4), pp. 342–4.

Koivunen, J, Aaltonen, V & Peltonen, J. 2006. Protein kinase C (PKC) family in cancer progression. *Cancer Lett* 235(1), pp. 1–10.

Kolaczkowska, E & Kubes, P. 2013. Neutrophil recruitment and function in health and inflammation. *Nat Rev Immunol* 13(3), pp. 159–75.

Kouzarides, T. 2007. Chromatin Modifications and Their Function. *Cell* 128(4), pp. 693–705.

Kovalchuk, O & Baulch, J. 2008. Epigenetic changes and nontargeted radiation effects–is there a link? *Environ Mol Mutagen* 49(1), pp. 16–25.

Kröber, A, Seiler, T, Benner, A, Bullinger, L, Brückle, E, Lichter, P, Döhner, H & Stilgenbauer, S. 2002. V(H) mutation status, CD38 expression level, genomic aberrations, and survival in chronic lymphocytic leukemia. *Blood* 100(4), pp. 1410–6.

Krysov, S, Dias, S, Paterson, A, Mockridge, CI, Potter, KN, Smith, KA, Ashton-Key, M, Stevenson, FK & Packham, G. 2012. Surface IgM stimulation induces MEK1/2-dependent MYC expression in chronic lymphocytic leukemia cells. *Cell* 119(1), pp. 170–9.

Kulis, M, Heath, S, Bibikova, M, Queirós, AC, Navarro, A, Clot, G, Martínez-Trillos, A, Castellano, G, Brun-Heath, I, Pinyol, M, Barberán-Soler, S, Papasaikas, P, Jares, P, Beà, S, Rico, D, Ecker, S, Rubio, M, Royo, R, Ho, V, Klotzle, B, Hernández, L, Conde, L, López-Guerra, M, Colomer, D, Villamor, N, Aymerich, M, Rozman, M, Bayes, M, Gut, M, Gelpí, JL, Orozco, M, Fan, JB, Quesada, V, Puente, XS, Pisano, DG, Valencia, A, López-Guillermo, A, Gut, I, López-Otín, C, Campo, E & Martín-Subero, JI. 2012. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet* 44(11), pp. 1236–42.

Kurotaki, N, Imaizumi, K, Harada, N, Masuno, M, Kondoh, T, Nagai, T, Ohashi, H, Naritomi, K, Tsukahara, M, Makita, Y, Sugimoto, T, Sonoda, T, Hasegawa, T, Chinen, Y, Tomita, HA, Kinoshita, A, Mizuguchi, T, Yoshiura, KI, Ohta, T, Kishino, T, Fukushima, Y, Niikawa, N & Matsumoto, N. 2002. Haploinsufficiency of NSD1 causes Sotos syndrome. *Nat Genet* 30(4), pp. 365–6.

Kutay, U, Ralf Bischoff, F, Kostka, S, Kraft, R & Görlich, D. 1997. Export of importin $\alpha$ from the nucleus is mediated by a specific nuclear transport factor. *Cell* 90(6), pp. 1061–71.

La Starza, R, Gorello, P, Rosati, R, Riezzo, A, Veronese, A, Ferrazzi, E, Martelli, MF, Negrini, M & Mecucci, C. 2004. Cryptic insertion producing two NUP98/NSD1 chimeric transcripts in adult refractory anemia with an excess of blasts. *Genes Chromosomes Cancer* 41(4), pp. 395–9.

Laird, PW. 2010. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* 11(3), pp. 191–203.

Lam, VK, Emberly, E, Fraser, HB, Neumann, SM, Chen, E, Miller, GE & Kobor, MS. 2012. Factors underlying variable DNA methylation in a human community cohort. *Proc Natl Acad Sci U S A* 109(Suppl 2), pp. 17253–60.

Lancichinetti, A & Fortunato, S. 2009. Community detection algorithms: A comparative analysis. *Phys Rev E Stat Nonlin Soft Matter Phys* 80(5), p. 056117.

Landau, DA, Carter, SL, Getz, G & Wu, CJ. 2014a. Clonal evolution in hematological malignancies and therapeutic implications. *Leukemia* 28(1), pp. 34–43.

Landau, DA, Carter, SL, Stojanov, P, McKenna, A, Stevenson, K, Lawrence, MS, Sougnez, C, Stewart, C, Sivachenko, A, Wang, L, Wan, Y, Zhang, W, Shukla, SA, Vartanov, A, Fernandes, SM, Saksena, G, Cibulskis, K, Tesar, B, Gabriel, S, Hacohen, N, Meyerson, M, Lander, ES, Neuberg, D, Brown, JR, Getz, G & Wu, CJ. 2013. Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell* 152(4), pp. 714–26.

Landau, DA, Clement, K, Ziller, MJ, Boyle, P, Fan, J, Gu, H, Stevenson, K, Sougnez, C, Wang, L, Li, S, Kotliar, D, Zhang, W, Ghandi, M, Garraway, L, Fernandes, SM, Livak, KJ, Gabriel, S, Gnirke, A, Lander, ES, Brown, JR, Neuberg, D, Kharchenko, PV, Hacohen, N, Getz, G, Meissner, A & Wu, CJ. 2014b. Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell* 26(6), pp. 813–25.

Law, CW, Chen, Y, Shi, W & Smyth, GK. 2014. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15(2), p. R29.

Lay, FD, Liu, Y, Kelly, TK, Witt, H, Farnham, PJ, Jones, PA & Berman, BP. 2015. The role of DNA methylation in directing the functional organization of the cancer epigenome. *Genome Res* 25(4), pp. 467–77.

Lee, CKK, Smith, E, Gimeno, R, Gertner, R & Levy, DE. 2000. STAT1 Affects Lymphocyte Survival and Proliferation Partially Independent of Its Role Downstream of IFN-y. *J Immunol* 164(3), pp. 1286–92.

Lee, L, Stollar, E, Chang, J, Gu, J, Brien, RO, Ladbury, J, Carpenter, B, Roberts, S & Luisi, B. 2001. Expression of the Oct-1 Transcription Factor and Characterization of Its Interactions. *Biochemistry* 40, pp. 6580–8.

Lefebvre, C, Rajbhandari, P, Alvarez, MJ, Bandaru, P, Lim, WK, Sato, M, Wang, K, Sumazin, P, Kustagi, M, Bisikirska, BC, Basso, K, Beltrao, P, Krogan, N, Gautier, J, Dalla-Favera, R & Califano, A. 2010. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol Syst Biol* 6(377).

Lehner, B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol Syst Biol* 4(170).

Lehner, B & Kaneko, K. 2011. Fluctuation and response in biology. *Cell Mol Life Sci* 68(6), pp. 1005–10.

Lengauer, C, Kinzler, KW & Vogelstein, B. 1998. Genetic instabilities in human cancers. *Nature* 396(6712), pp. 643–9.

Lev Maor, G, Yearim, A & Ast, G. 2015. The alternative role of DNA methylation in splicing regulation. *Trends Genet* 31(5), pp. 274–80.

Levene, H. 1960. Robust tests for equality of variances. *Contrib to Probab Stat. Essays Honor Harold Hotell* pp. 278–92.

Li, B, Carey, M & Workman, JL. 2007. The Role of Chromatin during Transcription. *Cell* 128(4), pp. 707–19.

Li, J, Liu, Y, Kim, T, Min, R & Zhang, Z. 2010. Gene expression variability within and between human populations and implications toward disease susceptibility. *PLoS Comput Biol* 6(8), p. e1000910.

Liaw, A & Wiener, M. 2002. Classification and Regression by randomForest. *R News* 2(3), pp. 18–22.

Lienert, F, Wirbelauer, C, Som, I, Dean, A, Mohn, F & Schübeler, D. 2011. Identification of genetic elements that autonomously determine DNA methylation states. *Nat Genet* 43(11), pp. 1091–7.

Loader, C. 1999. *Local Regression and Likelihood*. New York: Springer. Available at: http://www.jstor.org/stable/1270956?origin=crossref.

Lock, LF, Takagi, N & Martin, GR. 1987. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell* 48(1), pp. 39–46.

Longo, PG, Laurenti, L, Gobessi, S, Petlickovski, A, Pelosi, M, Chiusolo, P, Sica, S, Leone, G & Efremov, DG. 2007. The Akt signaling pathway determines the different proliferative capacity of chronic lymphocytic leukemia B-cells from patients with progressive and stable disease. *Leukemia* 21(1), pp. 110–20.

López-Otín, C, Blasco, MA, Partridge, L, Serrano, M & Kroemer, G. 2013. The Hallmarks of Aging. *Cell* 153(6), pp. 1194–217.

Love, MI, Huber, W & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol* 15(12), p. 550.

Lu, NZ, Wardell, SE, Burnstein, KL, Defranco, D, Fuller, PJ, Giguere, V, Hochberg, RB, McKay, L, Renoir, JM, Weigel, NL, Wilson, EM, McDonnell, DP & Cidlowski, JA. 2006. International Union of Pharmacology. LXV. The pharmacology and classification of the nuclear receptor superfamily: glucocorticoid, mineralocorticoid, progesterone, and androgen receptors. *Pharmacol Rev* 58(4), pp. 782–97.

Lubeck, E & Cai, L. 2012. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat Methods* 9(7), pp. 743–8.

Lucas, S, Ghilardi, N, Li, J & De Sauvage, FJ. 2003. IL-27 regulates IL-12 responsiveness of naive CD4+ T cells through Stat1-dependent and -independent mechanisms. *Proc Natl Acad Sci U S A* 100(25), pp. 15047–52.

Lucio-Eterovic, AK, Singh, MM, Gardner, JE, Veerappan, CS, Rice, JC & Carpenter, PB. 2010. Role for the nuclear receptor-binding SET domain protein 1 (NSD1) methyltransferase in coordinating lysine 36 methylation at histone 3 with RNA polymerase II function. *Proc Natl Acad Sci U S A* 107(39), pp. 16952–7.

Lüscher, B. 2001. Function and regulation of the transcription factors of the Myc/Max/Mad network. *Gene* 277(1-2), pp. 1–14.

Makismovic, J, Gordon, L & Oshlack, A. 2012. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol* 13(6), p. R44.

Mar, JC, Matigian, NA, Mackay-Sim, A, Mellick, GD, Sue, CM, Silburn, PA, McGrath, JJ, Quackenbush, J & Wells, CA. 2011. Variance of gene expression identifies altered network constraints in neurological disease. *PLoS Genet* 7(8), p. e1002207.

Markle, JG & Fish, EN. 2014. SeXX matters in immunity. *Trends Immunol* 35(3), pp. 97–104.

Martinez, NJ & Walhout, AJM. 2009. The interplay between transcription factors and microRNAs in genome-scale regulatory networks. *Bioessays* 31(4), pp. 435–45.

Marusyk, A, Almendro, V & Polyak, K. 2012. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* 12(5), pp. 323–34.

Mattick, JS, Amaral, PP, Dinger, ME, Mercer, TR & Mehler, MF. 2009. *RNA regulation of epigenetic processes*, [Online].

Maunakea, AK, Nagarajan, RP, Bilenky, M, Ballinger, TJ, D'Souza, C, Fouse, SD, Johnson, BE, Hong, C, Nielsen, C, Zhao, Y, Turecki, G, Delaney, A, Varhol, R, Thiessen, N, Shchors, K, Heine, VM, Rowitch, DH, Xing, X, Fiore, C, Schillebeeckx, M, Jones, SJM, Haussler, D, Marra, MA, Hirst, M, Wang, T & Costello, JF. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466(7303), pp. 253–7.

McAdams, HH & Arkin, A. 1997. Stochastic mechanisms in gene expression. *Proc Natl Acad Sci U S A* 94(3), pp. 814–9.

McCall, MN & Irizarry, RA. 2011. Thawing Frozen Robust Multi-array Analysis (fRMA). *BMC Bioinformatics* 12(1), p. 369.

Meixner, A, Karreth, F, Kenner, L & Wagner, EF. 2004. JunD regulates lymphocyte proliferation and T helper cell cytokine expression. *EMBO J* 23(6), pp. 1325–35.

Mencalha, AL, Binato, R, Ferreira, GM, Du Rocher, B & Abdelhay, E. 2012. Forkhead Box M1 (FoxM1) Gene Is a New STAT3 Transcriptional Factor Target and Is Essential for Proliferation, Survival and DNA Repair of K562 Cell Line. *PLoS One* 7(10), p. e48160.

Méndez-Ferrer, S, Lucas, D, Battista, M & Frenette, PS. 2008. Haematopoietic stem cell release is regulated by circadian oscillations. *Nature* 452(7186), pp. 442–7.

Messmer, BT, Albesiano, E, Messmer, D & Chiorazzi, N. 2012. The pattern and distribution of immunoglobulin V H gene mutations din chronic lymphocytic leukemia B cells are consistent with the canonical somatic hypermutation process. *Blood* 103(9), pp. 3490–5.

Messmer, BT, Messmer, D, Allen, SL, Kolitz, JE, Kudalkar, P, Cesar, D, Murphy, EJ, Koduru, P, Ferrarini, M, Zupo, S, Cutrona, G, Damle, RN, Wasil, T, Rai, KR, Hellerstein, MK & Chiorazzi, N. 2005. In vivo measurements document the dynamic cellular kinetics of chronic lymphocytic leukemia B cells. *J Clin Invest* 115(3), pp. 755–64.

Meunier, D, Lambiotte, R, Fornito, A, Ersche, KD & Bullmore, ET. 2009. Hierarchical modularity in human brain functional networks. *Front Neuroinform* 3, p. 37.

Miranti, CK & Brugge, JS. 2002. Sensing the environment: a historical perspective on integrin signal transduction. *Nat Cell Biol* 4(4), pp. E83–90.

Mócsai, A. 2013. Diverse novel functions of neutrophils in immunity, inflammation, and beyond. *J Exp Med* 210(7), pp. 1283–99.

Molloy, EJ, Neill, AJO, Grantham, JJ, Sheridan-Pereira, M, Fitzpatrick, JM, Webb, DW & Watson, RWG. 2013. Sex-specific alterations in neutrophil apoptosis: the role of estradiol and progesterone. *Phagocytes* 102(7), pp. 2653–9.

Montserrat, E, Fontanilles, M & Estape, J. 1991. Treatment of Chronic Lymphocytic Leukemia: A Preliminary Report of Spanish (Pethema) Trials. *Leuk Lymphoma* 5, pp. 89–91.

Morgan, HD, Sutherland, HG, Martin, DI & Whitelaw, E. 1999. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* 23(3), pp. 314–8.

Morley, M, Molony, CM, Weber, TM, Devlin, JL, Ewens, KG, Spielman, RS & Cheung, VG. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature* 430(7001), pp. 743–7.

Mraz, M, Pospisilova, S, Malinova, K, Slapak, I & Mayer, J. 2009. MicroRNAs in chronic lymphocytic leukemia pathogenesis and disease subtypes. *Leuk Lymphoma* 50(3), pp. 506–9.

Na, SY, Choi, JE, Kim, HJ, Jhun, BH, Lee, YC & Lee, JW. 1999. Bcl3, an I$\kappa$B Protein, Stimulates Activating Protein-1 Transactivation and Cellular Proliferation. *J Biol Chem* 274(40), pp. 28491–6.

Naranbhai, V, Fairfax, BP, Makino, S, Humburg, P, Wong, D, Ng, E, Hill, AVS & Knight, JC. 2015. Genomic modulators of gene expression in human neutrophils. *Nat Comm* 6, p. 7545.

Neefjes, J, Jongsma, MLM, Paul, P & Bakke, O. 2011. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* 11(12), pp. 823–36.

Neubauer, O, Sabapathy, S, Lazarus, R, Jowett, JBM, Desbrow, B, Peake, JM, Cameron-Smith, D, Haseler, LJ, Wagner, KH & Bulmer, AC. 2013. Transcriptome analysis of

neutrophils after endurance exercise reveals novel signaling mechanisms in the immune response to physiological stress. *J Appl Physiol* 114(12), pp. 1677–88.

Neufert, C, Becker, C, Wirtz, S, Fantini, MC, Weigmann, B, Galle, PR & Neurath, MF. 2007. IL-27 controls the development of inducible regulatory T cells and Th17 cells via differential effects on STAT1. *Eur J Immunol* 37(7), pp. 1809–16.

Newman, MEJ. 2012. Communities, modules and large-scale structure in networks. *Nat Phys* 8, pp. 25–31.

Nicholson, S, Whitehouse, H, Naidoo, K & Byers, RJ. 2011. Yin Yang 1 in human cancer. *Crit Rev Oncog* 16(3-4), pp. 245–60.

Nilsson, JA & Cleveland, JL. 2003. Myc pathways provoking cell suicide and cancer. *Oncogene* 22(56), pp. 9007–21.

Oakes, CC, Claus, R, Gu, L, Assenov, Y, Hüllein, J, Zucknick, M, Bieg, M, Brocks, D, Bogatyrova, O, Schmidt, CR, Rassenti, L, Kipps, TJ, Mertens, D, Lichter, P, Döhner, H, Stilgenbauer, S, Byrd, JC, Zenz, T & Plass, C. 2014. Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia. *Cancer Discov* 4(3), pp. 348–61.

Obermann, EC, Went, P, Tzankov, A, Pileri, SA, Hofstaedter, F, Marienhagen, J, Stoehr, R & Dirnhofer, S. 2007. Cell cycle phase distribution analysis in chronic lymphocytic leukaemia: a significant number of cells reside in early G1-phase. *J Clin Pathol* 60(7), pp. 794–7.

Odintsova, TI, Müller, EC, Ivanov, AV, Egorov, TA, Bienert, R, Vladimirov, SN, Kostka, S, Otto, A, Wittmann-Liebold, B & Karpova, GG. 2003. Characterization and analysis of posttranslational modifications of the human large cytoplasmic ribosomal subunit proteins by mass spectrometry and Edman sequencing. *J Protein Chem* 22(3), pp. 249–58.

Ohm, JE, McGarvey, KM, Yu, X, Cheng, L, Schuebel, KE, Cope, L, Mohammad, HP, Chen, W, Daniel, VC, Yu, W, Berman, DM, Jenuwein, T, Pruitt, K, Sharkis, SJ, Watkins, DN, Herman, JG & Baylin, SB. 2007. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet* 39(2), pp. 237–42.

OpenStax College. 2013. *The Hematopoietic System of the Bone Marrow*, [Online]. Available at: http://cnx.org/content/col11496/1.6/. Accessed 2013-12-09.

Osaki, E, Nishina, Y, Inazawa, J, Copeland, NG, Gilbert, DJ, Jenkins, NA, Ohsugi, M, Tezuka, T, Yoshida, M & Semba, K. 1999. Identification of a novel Sry-related gene and its germ cell-specific expression. *Nucleic Acids Res* 27(12), pp. 2503–10.

Ozbudak, EM, Thattai, M, Kurtser, I, Grossman, AD & van Oudenaarden, A. 2002. Regulation of noise in the expression of a single gene. *Nat Genet* 31(1), pp. 69–73.

Paszek, P, Ryan, S, Ashall, L, Sillitoe, K, Harper, CV, Spiller, DG, Rand, DA & White, MRH. 2010. Population robustness arising from cellular heterogeneity. *Proc Natl Acad Sci U S A* 107(25), pp. 11644–9.

Pede, V, Rombout, A, Vermeire, J, Naessens, E, Mestdagh, P, Robberecht, N, Vanderstraeten, H, Van Roy, N, Vandesompele, J, Speleman, F, Philippé, J & Verhasselt, B. 2013. CLL cells respond to B-Cell receptor stimulation with a microRNA/mRNA signature associated with MYC activation and cell cycle progression. *PLoS One* 8(4), p. e60275.

Pelengaris, S, Khan, M & Evan, G. 2002. c-MYC: more than just a matter of life and death. *Nat Rev Cancer* 2(10), pp. 764–76.

Pezzella, F, Tse, AG, Cordell, JL, Pulford, KA, Gatter, KC & Mason, DY. 1990. Expression of the bcl-2 oncogene protein is not specific for the 14;18 chromosomal translocation. *Am J Pathol* 137(2), pp. 225–32.

Pflanz, S, Timans, JC, Cheung, J, Rosales, R, Kanzler, H, Gilbert, J, Hibbert, L, Churakova, T, Travis, M, Vaisberg, E, Blumenschein, WM, Mattson, JD, Wagner, JL, To, W, Zurawski, S, McClanahan, TK, Gorman, DM, Bazan, JF, De Waal Malefyt, R, Rennick, D & Kastelein, RA. 2002. IL-27, a heterodimeric cytokine composed of EBI3 and p28 protein, induces proliferation of naive CD4+ T cells. *Immunity* 16(6), pp. 779–90.

Phipson, B & Oshlack, A. 2014. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biol* 15(9), p. 465.

Pillay, J, Braber, ID, Vrisekoop, N, Kwast, LM, Boer, RJD, Borghans, AM, Tesselaar, K & Koenderman, L. 2010. Brief report In vivo labeling with 2H2O reveals a human neutrophil lifespan of 5.4 days. *Blood* 116(4), pp. 625–7.

Polakis, P. 2000. Wnt signaling and cancer. *Genes Dev* 14(15), pp. 1837–51.

Powell, ND, Sloan, EK, Bailey, MT, Arevalo, JMG, Miller, GE, Chen, E, Kobor, MS, Reader, BF, Sheridan, JF & Cole, SW. 2013. Social stress up-regulates inflammatory gene expression in the leukocyte transcriptome via $\beta$-adrenergic induction of myelopoiesis. *Proc Natl Acad Sci U S A* 110(41), pp. 16574–9.

Prieto, C, Rivas, MJ, Sánchez, JM, López-Fidalgo, J & De Las Rivas, J. 2006. Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. *Bioinformatics* 22(9), pp. 1103–10.

Puente, XS, Pinyol, M, Quesada, V, Conde, L, Ordóñez, GR, Villamor, N, Escaramis, G, Jares, P, Beà, S, González-Díaz, M, Bassaganyas, L, Baumann, T, Juan, M, López-Guerra, M, Colomer, D, Tubío, JMC, López, C, Navarro, A, Tornador, C, Aymerich, M, Rozman, M, Hernández, JM, Puente, DA, Freije, JMP, Velasco, G, Gutiérrez-Fernández, A, Costa, D, Carrió, A, Guijarro, S, Enjuanes, A, Hernández, L, Yagüe, J, Nicolás, P, Romeo-Casabona, CM, Himmelbauer, H, Castillo, E, Dohm, JC, de Sanjosé, S, Piris, MA, de Alava, E, Miguel, JS, Royo, R, Gelpí, JL, Torrents, D, Orozco, M, Pisano, DG, Valencia, A, Guigó, R, Bayés, M, Heath, S, Gut, M, Klatt, P, Marshall, J, Raine, K, Stebbings, LA, Futreal, PA, Stratton, MR, Campbell, PJ, Gut, I, López-Guillermo, A, Estivill, X, Montserrat, E, López-Otín, C & Campo, E. 2011. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* 475(7354), pp. 101–5.

Pujadas, E & Feinberg, AP. 2012. Regulated noise in the epigenetic landscape of development and disease. *Cell* 148(6), pp. 1123–31.

Pyrpasopoulou, A, Meier, J, Maison, C, Simos, G & Georgatos, SD. 1996. The lamin B receptor (LBR) provides essential chromatin docking sites at the nuclear envelope. *EMBO J* 15(24), pp. 7108–19.

Qiao, Q, Li, Y, Chen, Z, Wang, M, Reinberg, D & Xu, RM. 2011. The structure of NSD1 reveals an autoregulatory mechanism underlying histone H3K36 methylation. *J Biol Chem* 286(10), pp. 8361–8.

Qu, GZ, Grundy, PE, Narayan, A & Ehrlich, M. 1999. Frequent hypomethylation in Wilms tumors of pericentromeric DNA in chromosomes 1 and 16. *Cancer Genet Cytogenet* 109(1), pp. 34–9.

Queirós, AC, Villamor, N, Clot, G, Martinez-Trillos, A, Kulis, M, Navarro, A, Penas, EMM, Jayne, S, Majid, A, Richter, J, Bergmann, AK, Kolarova, J, Royo, C, Russiñol, N, Castellano, G, Pinyol, M, Bea, S, Salaverria, I, López-Guerra, M, Colomer, D,

Aymerich, M, Rozman, M, Delgado, J, Giné, E, González-Díaz, M, Puente, XS, Siebert, R, Dyer, MJS, López-Otín, C, Rozman, C, Campo, E, López-Guillermo, A & Martín-Subero, JI. 2014. A B-cell epigenetic signature defines three biological subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia* 29(3), pp. 598–605.

Quesada, V, Conde, L, Villamor, N, Ordóñez, GR, Jares, P, Bassaganyas, L, Ramsay, AJ, Beà, S, Pinyol, M, Martínez-Trillos, A, López-Guerra, M, Colomer, D, Navarro, A, Baumann, T, Aymerich, M, Rozman, M, Delgado, J, Giné, E, Hernández, JM, González-Díaz, M, Puente, DA, Velasco, G, Freije, JMP, Tubío, JMC, Royo, R, Gelpí, JL, Orozco, M, Pisano, DG, Zamora, J, Vázquez, M, Valencia, A, Himmelbauer, H, Bayés, M, Heath, S, Gut, M, Gut, I, Estivill, X, López-Guillermo, A, Puente, XS, Campo, E & López-Otín, C. 2011. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* 44(1), pp. 47–52.

Quiroga, M, Balakrishnan, K, Kurtova, AV, Sivina, M, Keating, MJ, Wierda, WG, Gandhi, V & Burger, JA. 2009. B cell antigen receptor signaling enhances chronic lymphocytic leukemia cell migration and survival: specific targeting with a novel spleen tyrosine kinase inhibitor, R406. *Blood* 114(5), pp. 1029–37.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. Vienna, Available at: http://www.r-project.org.

Radom-Aizik, S, Zaldivar, FJ, Leu, S, Galassetti, P & Cooper, D. 2008. Effects of 30 min of aerobic exercise on gene expression in human neutrophils. *J Appl Physiol* 104(1), pp. 236–43.

Rai, KR & Han, T. 1990. Prognostic factors and clinical staging in chronic lymphocytic leukemia. *Hematol Oncol Clin North Am* 4(2), p. 447.

Rai, KR, Sawitsky, A, Cronkite, EP, Chanana, AD, Levy, RN & Pasternack, BS. 1975. Clinical Staging of Chronic Lymphocytic Leukemia. *Blood* 46(2), pp. 219–34.

Raj, A, Peskin, CS, Tranchina, D, Vargas, DY & Tyagi, S. 2006. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4(10), p. e309.

Raj, A, Rifkin, SA, Andersen, E & Van Oudenaarden, A. 2010. Variability in gene expression underlies incomplete penetrance. *Nature* 463(7283), pp. 913–8.

Raj, A & Van Oudenaarden, A. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135(2), pp. 216–26.

Rakyan, VK, Down, TA, Maslau, S, Andrew, T, Yang, TP, Beyan, H, Whittaker, P, McCann, OT, Finer, S, Valdes, AM, Leslie, RD, Deloukas, P & Spector, TD. 2010. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res* 20(4), pp. 434–9.

Rampazzo, E, Bonaldi, L, Trentin, L, Visco, C, Keppel, S, Giunco, S, Frezzato, F, Facco, M, Novella, E, Giaretta, I, Del Bianco, P, Semenzato, G & De Rossi, A. 2012. Telomere length and telomerase levels delineate subgroups of B-cell chronic lymphocytic leukemia with different biological characteristics and clinical outcomes. *Haematologica* 97(1), pp. 56–63.

Ramsay, AJ, Quesada, V, Foronda, M, Conde, L, Martínez-Trillos, A, Villamor, N, Rodríguez, D, Kwarciak, A, Garabaya, C, Gallardo, M, López-Guerra, M, López-Guillermo, A, Puente, XS, Blasco, MA, Campo, E & López-Otín, C. 2013. POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat Genet* 45(5), pp. 526–30.

Rana, S, Munawar, M, Shahid, A, Malik, M, Ullah, H, Fatima, W, Mohsin, S & Mahmood, S. 2014. Deregulated expression of circadian clock and clock-controlled cell cycle genes in chronic lymphocytic leukemia. *Mol Biol Rep* 41(1), pp. 95–103.

Rapaport, F, Khanin, R, Liang, Y, Pirun, M, Krek, A, Zumbo, P, Mason, CE, Socci, ND & Betel, D. 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 14(9), p. R95.

Rasala, BA, Orjalo, AV, Shen, Z, Briggs, S & Forbes, DJ. 2006. ELYS is a dual nucleoporin/kinetochore protein required for nuclear pore assembly and proper cell division. *Proc Natl Acad Sci U S A* 103(47), pp. 17801–6.

Raser, JM & O'Shea, EK. 2005. Noise in Gene Expression: Origins, Consequences, and Control. *Science* 309(5743), pp. 2010–3.

Rassenti, LZ, Jain, S, Keating, MJ, Wierda, WG, Grever, MR, Byrd, JC, Kay, NE, Brown, JR, Gribben, JG, Neuberg, DS, He, F, Greaves, AW, Rai, KR & Kipps, TJ. 2008. Relative value of ZAP-70, CD38, and immunoglobulin mutation status in predicting aggressive disease in chronic lymphocytic leukemia. *Blood* 112(5), pp. 1923–30.

Reik, W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447(7143), pp. 425–32.

Reya, T, Morrison, SJ, Clarke, MF & Weissman, IL. 2001. Stem cells, cancer, and cancer stem cells. *Nature* 414(6859), pp. 105–11.

Reya, T, O'Riordan, M, Okamura, R, Devaney, E, Willert, K, Nusse, R & Grosschedl, R. 2000. Wnt signaling regulates B lymphocyte proliferation through a LEF-1 dependent mechanism. *Immunity* 13(1), pp. 15–24.

Richards, EJ. 2006. Inherited epigenetic variation–revisiting soft inheritance. *Nat Rev Genet* 7(5), pp. 395–401.

Riggs, AD. 1975. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet* 14(1), pp. 9–25.

Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W & Smyth, GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7), p. e47.

Rivas, MA, Pirinen, M, Conrad, DF, Lek, M, Tsang, EK, Karczewski, KJ, Maller, JB, Kukurba, KR, Deluca, DS, Fromer, M, Ferreira, PG, Smith, KS, Zhang, R, Zhao, F, Banks, E, Poplin, R, Ruderfer, DM, Purcell, SM, Tukiainen, T, Minikel, EV, Stenson, PD, Cooper, DN, Huang, KH, Sullivan, TJ, Nedzel, J, GTEx Consortium, Bustamante, CD, Li, JB, Daly, MJ, Guigo, R, Donnelly, P, Ardlie, K, Sammeth, M, Dermitzakis, ET, Mccarthy, MI & Montgomery, SB. 2015. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* 348(6235), pp. 666–9.

Roadmap Epigenomics Consortium. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539), pp. 317–30.

Roberts, SB, Segil, N & Heintz, N. 1991. Differential phosphorylation of the transcription factor Oct1 during the cell cycle. *Science* 253(5023), pp. 1022–6.

Robinson, MD, McCarthy, DJ & Smyth, GK. 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1), pp. 139–40.

Roche, PA & Furuta, K. 2015. The ins and outs of MHC class II-mediated antigen processing and presentation. *Nat Rev Immunol* 15(4), pp. 203–16.

Rockman, MV & Kruglyak, L. 2006. Genetics of global gene expression. *Nat Rev Genet* 7(11), pp. 862–72.

Rodriguez, J, Frigola, J, Vendrell, E, Risques, RA, Fraga, MF, Morales, C, Moreno, V, Esteller, M, Capellà, G, Ribas, M & Peinado, MA. 2006. Chromosomal instability

correlates with genome-wide DNA demethylation in human primary colorectal cancers. *Cancer Res* 66(17), pp. 8462–9468.

Roos, G, Kröber, A, Grabowski, P, Kienle, D, Bühler, A, Döhner, H, Rosenquist, R & Stilgenbauer, S. 2008. Short telomeres are associated with genetic complexity, high-risk genomic aberrations, and short survival in chronic lymphocytic leukemia. *Blood* 111(4), pp. 2246–52.

Rosenwald, A, Alizadeh, AA, Widhopf, G, Simon, R, Davis, RE, Yu, X, Yang, L, Pickeral, OK, Rassenti, LZ, Powell, J, Botstein, D, Byrd, JC, Grever, MR, Cheson, BD, Chiorazzi, N, Wilson, WH, Kipps, TJ, Brown, PO & Staudt, LM. 2001. Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia. *J Exp Med* 194(11), pp. 1639–47.

Rozman, C & Montserrat, E. 1995. Chronic Lymphocytic Leukemia. *N Engl J Med* 333(16), pp. 1052–57.

Ruvolo, PP, Deng, X, Carr, BK & May, WS. 1998. A Functional Role for Mitochondrial Protein Kinase C in Bcl2 Phosphorylation and Suppression of Apoptosis. *J Biol Chem* 273(39), pp. 25436–42.

Sanchez-Mut, JV, Aso, E, Heyn, H, Matsuda, T, Bock, C, Ferrer, I & Esteller, M. 2014. Promoter hypermethylation of the phosphatase DUSP22 mediates PKA-dependent TAU phosphorylation and CREB activation in Alzheimer's disease. *Hippocampus* 24(4), pp. 363–8.

Sardet, C, Vidal, M, Cobrinik, D, Geng, Y, Onufryk, C, Chen, A & Weinberg, RA. 1995. E2F-4 and E2F-5, two members of the E2F family, are expressed in the early phases of the cell cycle. *Proc Natl Acad Sci U S A* 92(6), pp. 2403–7.

Satija, R & Shalek, AK. 2014. Heterogeneity in immune responses: from populations to single cells. *Trends Immunol* 35(5), pp. 219–29.

Scheiermann, C, Kunisaki, Y, Lucas, D, Chow, A, Jang, JE, Zhang, D, Hashimoto, D, Merad, M & Frenette, PS. 2012. Adrenergic nerves govern circadian leukocyte recruitment to tissues. *Immunity* 37(2), pp. 290–301.

Schena, M. 2003. *Microarray Analysis*. Hoboken, New Jersey: John Wiley & Sons, 1 ed.

Schlesinger, Y, Straussman, R, Keshet, I, Farkash, S, Hecht, M, Zimmerman, J, Eden, E, Yakhini, Z, Ben-Shushan, E, Reubinoff, BE, Bergman, Y, Simon, I & Cedar, H. 2007.

Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat Genet* 39(2), pp. 232–6.

Schuh, A, Becq, J, Humphray, S, Alexa, A, Burns, A, Clifford, R, Feller, SM, Grocock, R, Henderson, S, Khrebtukova, I, Kingsbury, Z, Luo, S, McBride, D, Murray, L, Menju, T, Timbs, A, Ross, M, Taylor, J & Bentley, D. 2012. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* 120(20), pp. 4191–6.

Schuurs, AHWM & Verheul, HAM. 1990. Effects of gender and sex steroids immune response. *J Steroid Biochem* 35(2), pp. 157–172.

Scotland, RS, Stables, MJ, Madalli, S, Watson, P & Gilroy, DW. 2011. Sex differences in resident immune cell phenotype underlie more efficient acute inflammatory responses in female mice. *Blood* 118(22), pp. 5918–27.

Segil, N, Roberts, SB & Heintz, N. 1991. Mitotic phosphorylation of the Oct-1 homeodomain and regulation of Oct-1 DNA binding activity. *Science* 254(5039), pp. 1814–6.

Sekine, Y, Ikeda, O, Hayakawa, Y, Tsuji, S, Imoto, S, Aoki, N, Sugiyama, K & Matsuda, T. 2007. DUSP22/LMW-DSP2 regulates estrogen receptor-alpha-mediated signaling through dephosphorylation of Ser-118. *Oncogene* 26(41), pp. 6038–49.

Seligmann, B, Chused, TM & Gallin, JI. 1981. Human neutrophil heterogeneity identified using flow microfluorometry to monitor membrane potential. *J Clin Invest* 68(5), pp. 1125–31.

Sellmann, L, De Beer, D, Bartels, M, Opalka, B, Nückel, H, Dührsen, U, Dürig, J, Seifert, M, Siemer, D, Küppers, R, Baerlocher, GM & Röth, A. 2011. Telomeres and prognosis in patients with chronic lymphocytic leukaemia. *Int J Hematol* 93(1), pp. 74–82.

Seyednasrollah, F, Laiho, A & Elo, LL. 2015. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform* 16(1), pp. 59–70.

Shain, KH & Tao, J. 2013. The B-cell receptor orchestrates environment-mediated lymphoma survival and drug resistance in B-cell malignancies. *Oncogene* 33(32), pp. 4107–13.

Sharma, S, Kelly, TK & Jones, PA. 2009. Epigenetics in cancer. *Carcinogenesis* 31(1), pp. 27–36.

Sharma, SV, Lee, DY, Li, B, Quinlan, MP, Takahashi, F, Maheswaran, S, McDermott, U, Azizian, N, Zou, L, Fischbach, MA, Wong, KK, Brandstetter, K, Wittner, B, Ramaswamy, S, Classon, M & Settleman, J. 2010. A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. *Cell* 141(1), pp. 69–80.

Shatz, M & Liscovitch, M. 2004. Caveolin-1 and cancer multidrug resistance: coordinate regulation of pro-survival proteins? *Leuk Res* 28(9), pp. 907–8.

Shen, Y, Luche, R, Wei, B, Gordon, ML, Diltz, CD & Tonks, NK. 2001. Activation of the Jnk signaling pathway by a dual-specificity phosphatase, JSP-1. *Proc Natl Acad Sci U S A* 98(24), pp. 13613–8.

Shupnik, MA. 2004. Crosstalk between steroid receptors and the c-Src-receptor tyrosine kinase pathways: implications for cell proliferation. *Oncogene* 23(48), pp. 7979–89.

Sing, T, Sander, O, Beerenwinkel, N & Lengauer, T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics* 21(20), p. 7881.

Smith, AK, Kilaru, V, Kocak, M, Almli, LM, Mercer, KB, Ressler, KJ, Tylavsky, FA & Conneely, KN. 2014. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics* 15(1), p. 145.

Smith, JA. 1994. Neutrophils, host defense, and inflammation: a double-edged sword. *J Leukoc Biol* 56(6), pp. 672–86.

Smyth, GK. 2005. Limma: Linear Models for Microarray Data. In: Gentleman, R, Carey, V, Dudoit, S, Irizarry, R & Huber, W, eds., *Bioinformatics and computational biology solutions using R and Bioconductor*, New York: Springer, pp. 397–420.

Snedecor, GW & Cochran, WG. 1989. *Statistical Methods*. Ames: Iwoa State University Press, 8 ed.

Snijder, B & Pelkmans, L. 2011. Origins of regulated cell-to-cell variability. *Nat Rev Mol Cell Biol* 12(2), pp. 119–25.

Somel, M, Khaitovich, P, Bahn, S, Pääbo, S & Lachmann, M. 2006. Gene expression becomes heterogeneous with age. *Curr Biol* 16(10), pp. R359–60.

Soneson, C & Delorenzi, M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14(1), p. 91.

Southworth, LK, Owen, AB & Kim, SK. 2009. Aging mice show a decreasing correlation of gene expression within genetic modules. *PLoS Genet* 5(12), p. e1000776.

Spencer, SL, Gaudet, S, Albeck, JG, Burke, JM & Sorger, PK. 2009. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* 459(7245), pp. 428–32.

Spitzer, JA & Zhang, P. 1996. Gender differences in neutrophil function and cytokine-induced neutrophil chemoattractant generation in endotoxic rats. *Inflammation* 20(5), pp. 485–98.

Stanley, ER, Berg, KL, Einstein, DB, Lee, PSW, Pixley, FJ, Wang, Y & Yeung, YG. 1997. Biology and action of colony-stimulating factor-1. *Mol Reprod Dev* 46(1), pp. 4–10.

Stemcell Technologies. 2015. Frequencies of Cell Types. Chart, Stemcell Technologies, Available at: http://www.stemcell.com/~/media/Files/wallchart_CellTypes_WEB.pdf.

Stilgenbauer, S, Sander, S, Bullinger, L, Benner, A, Leupolt, E, Winkler, D, Kröber, A, Kienle, D, Lichter, P & Döhner, H. 2007. Clonal evolution in chronic lymphocytic leukemia: Acquisition of high-risk genomic aberrations associated with unmutated VH, resistance to therapy, and short survival. *Haematologica* 92(9), pp. 1242–5.

Strahl, BD & Allis, CD. 2000. The language of covalent histone modifications. *Nature* 403(6765), pp. 41–5.

Stranger, BE, Nica, AC, Forrest, MS, Dimas, A, Bird, CP, Beazley, C, Ingle, CE, Dunning, M, Flicek, P, Koller, D, Montgomery, S, Tavaré, S, Deloukas, P & Dermitzakis, ET. 2007. Population genomics of human gene expression. *Nat Genet* 39(10), pp. 1217–24.

Subrahmanyam, YVBK, Yamaga, S, Prashar, Y, Lee, HH, Hoe, NP, Kluger, Y, Gerstein, M, Goguen, JD, Newburger, PE & Weissman, SM. 2001. RNA expression patterns change dramatically in human neutrophils exposed to bacteria. *Blood* 97(8), pp. 2457–68.

Sui, G. 2009. The Regulation of YY1 in Tumorigenesis and its Targeting Potential in Cancer Therapy. *Mol Cell Pharmacol* 1(3), pp. 157–76.

Summers, C, Rankin, SM, Condliffe, AM, Singh, N, Peters, AM & Chilvers, ER. 2010. Neutrophil kinetics in health and disease. *Trends Immunol* 31(8), pp. 318–24.

Suter, DM, Molina, N, Gatfield, D, Schneider, K, Schibler, U & Naef, F. 2011. Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science* 332(6028), pp. 472–4.

Swain, PS, Elowitz, MB & Siggia, ED. 2002. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci U S A* 99(20), pp. 12795–800.

Swanton, C & Beck, S. 2014. Epigenetic Noise Fuels Cancer Evolution. *Cancer Cell* 26(6), pp. 775–6.

Takashima, A & Yao, Y. 2015. Neutrophil plasticity: acquisition of phenotype and functionality of antigen-presenting cell. *J Leukoc Biol* pp. jlb–1MR1014.

Tanaka, N, Nakamura, E, Ohkura, M, Kuwabara, M, Yamashita, A, Onitsuka, T, Asada, Y, Hisa, H & Yamamoto, R. 2008. Both 5-hydroxytryptamine 5-HT2A and 5-HT1B receptors are involved in the vasoconstrictor response to 5-HT in the human isolated internal thoracic artery. *Clin Exp Pharmacol Physiol* 35(7), pp. 836–40.

Teschendorff, AE, Menon, U, Gentry-Maharaj, A, Ramus, SJ, Weisenberger, DJ, Shen, H, Campan, M, Noushmehr, H, Bell, CG, Maxwell, AP, Savage, DA, Mueller-Holzner, E, Marth, C, Kocjan, G, Gayther, SA, Jones, A, Beck, S, Wagner, W, Laird, PW, Jacobs, IJ & Widschwendter, M. 2010. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res* 20(4), pp. 440–6.

Teschendorff, AE & Widschwendter, M. 2012. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics* 28(11), pp. 1487–94.

Thompson, MA. 2006. *Chronic lymphocytic leukemia*, [Online]. Available at: http://en.wikipedia.org/wiki/File:Chronic_lymphocytic_leukemia.jpg.

Tillack, K, Naegele, M, Haueis, C, Schippling, S, Wandinger, KP, Martin, R & Sospedra, M. 2013. Gender differences in circulating levels of neutrophil extracellular traps in serum of multiple sclerosis patients. *J Neuroimmunol* 261(1-2), pp. 108–19.

Timp, W & Feinberg, AP. 2013. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat Rev Cancer* 13(7), pp. 497–510.

Tirosh, I, Reikhav, S, Levy, AA & Barkai, N. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* 324(5927), pp. 659–62.

Tong, WG, Wierda, WG, Lin, E, Kuang, SQ, Bekele, BN, Estrov, Z, Wei, Y, Yang, H, Keating, MJ & Garcia-Manero, G. 2010. Genome-wide DNA methylation profiling of chronic lymphocytic leukemia allows identification of epigenetically repressed molecular pathways with clinical impact. *Epigenetics* 5(6), pp. 499–508.

Tsukamoto, H, Clise-Dwyer, K, Huston, GE, Duso, DK, Buck, AL, Johnson, LL, Haynes, L & Swain, SL. 2009. Age-associated increase in lifespan of naive CD4 T cells contributes to T-cell homeostasis but facilitates development of functional defects. *Proc Natl Acad Sci U S A* 106(43), pp. 18333–8.

Tu, Y, Stolovitzky, G & Klein, U. 2002. Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci U S A* 99(22), pp. 14031–6.

Turro, E, Su, SY, Gonçalves, A, Coin, LJ, Richardson, S & Lewin, A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* 12(2), p. R13.

Tye, BK. 1999. MCM proteins in DNA replication. *Annu Rev Biochem* 68(1), pp. 649–86.

Vallat, L, Magdele, H, Kruhoffer, M, Sabatier, L, Orntoft, TF & Delic, J. 2003. The resistance of B-CLL cells to DNA damage-induced apoptosis defined by DNA microarrays. *Blood* 101(11), pp. 4598–606.

Van de Kar, LD, Javed, A, Zhang, Y, Serres, F, Raap, DK & Gray, TS. 2001. 5-HT2A receptors stimulate ACTH, corticosterone, oxytocin, renin, and prolactin release and activate hypothalamic CRF and oxytocin-expressing cells. *J Neurosci* 21(10), pp. 3572–9.

Veenendaal, MVE, Painter, RC, De Rooij, SR, Bossuyt, PMM, Van Der Post, JAM, Gluckman, PD, Hanson, MA & Roseboom, TJ. 2013. Transgenerational effects of prenatal exposure to the 1944-45 Dutch famine. *BJOG* 120(5), pp. 548–53.

Voisin, S, Eynon, N, Yan, X & Bishop, DJ. 2015. Exercise training and DNA methylation in humans. *Acta Physiol* 213(1), pp. 39–59.

Voronina, EN, Kolokoltsova, TD, Slinko, NM, Nechaeva, EA & Filipenko, ML. 2008. Transcription factor YY1 is involved in activation of transcription of the human gene for ribosomal protein L11. *Mol Biol* 42(1), pp. 98–104.

Voss, TC & Hager, GL. 2014. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat Rev Genet* 15(2), pp. 69–81.

Waddington, C. 1939. *An Introduction to Modern Genetics.* New York: Macmillan.

Waddington, C. 1942. The epigenotype. *Endeavour* 1, pp. 18–20.

Wagner, JR, Busche, S, Ge, B, Kwan, T, Pastinen, T & Blanchette, M. 2014. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol* 15(2), p. R37.

Wang, GG, Cai, L, Pasillas, MP & Kamps, MP. 2007. NUP98-NSD1 links H3K36 methylation to Hox-A gene activation and leukaemogenesis. *Nat Cell Biol* 9(7), pp. 804–12.

Wang, L, Lawrence, MS, Wan, Y, Stojanov, P, Sougnez, C, Stevenson, K, Werner, L, Sivachenko, A, DeLuca, DS, Zhang, L, Zhang, W, Vartanov, AR, Fernandes, SM, Goldstein, NR, Folco, EG, Cibulskis, K, Tesar, B, Sievers, QL, Shefler, E, Gabriel, S, Hacohen, N, Reed, R, Meyerson, M, Golub, TR, Lander, ES, Neuberg, D, Brown, JR, Getz, G & Wu, CJ. 2011. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med* 365(26), pp. 2497–506.

Wang, Q & Zhou, T. 2014. Alternative-splicing-mediated gene expression. *Phys Rev E Stat Nonlin Soft Matter Phys* 89(1), p. 12713.

Warren, AJ, Colledge, WH, Carlton, MBL, Evans, MJ, Smith, AJH & Rabbitts, TH. 1994. The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development. *Cell* 78(1), pp. 45–57.

Weitzman, JB, Fiette, L, Matsuo, K & Yaniv, M. 2000. JunD protects cells from p53-dependent senescence and apoptosis. *Mol Cell* 6(5), pp. 1109–19.

Wheater, PR, Burkitt, HG, Daniels, VG & Others. 1979. *Functional histology. A text and colour atlas.* Edinburgh: Churchill Livingstone.

Widschwendter, M, Fiegl, H, Egle, D, Mueller-Holzner, E, Spizzo, G, Marth, C, Weisenberger, DJ, Campan, M, Young, J, Jacobs, I & Laird, PW. 2007. Epigenetic stem cell signature in cancer. *Nat Genet* 39(2), pp. 157–8.

Wierstra, I & Alves, J. 2007. FOXM1, a typical proliferation-associated transcription factor. *Biol Chem* 388(12), pp. 1257–74.

Williams, GV, Rao, SG & Goldman-Rakic, PS. 2002. The physiological role of 5-HT2A receptors in working memory. *J Neurosci* 22(7), pp. 2843–54.

Wirths, S, Bugl, S & Kopp, HG. 2014. Neutrophil homeostasis and its regulation by danger signaling. *Blood* 123(23), pp. 3563–6.

Witte, T, Plass, C & Gerhauser, C. 2014. Pan-cancer patterns of DNA methylation. *Genome Med* 6(66), pp. 1–18.

Wolter, S, Doerrie, A, Weber, A, Schneider, H, Hoffmann, E, von der Ohe, J, Bakiri, L, Wagner, EF, Resch, K & Kracht, M. 2008. c-Jun controls histone modifications, NF-kappaB recruitment, and RNA polymerase II function to activate the ccl2 gene. *Mol Cell Biol* 28(13), pp. 4407–23.

Wong, JJ, Ritchie, W, Ebner, OA, Selbach, M, Wong, JWH, Huang, Y, Gao, D, Pinello, N, Gonzalez, M, Baidya, K, Thoeng, A, Khoo, TL, Bailey, CG, Holst, J & Rasko, JEJ. 2013. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* 154(3), pp. 583–95.

Wright, HL, Moots, RJ, Bucknall, RC & Edwards, SW. 2010. Neutrophil function in inflammation and inflammatory diseases. *Rheumatology* 49(9), pp. 1618–31.

Wu, TD & Nacu, S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26(7), pp. 873–81.

Xu, LL, Warren, MK, Rose, WL, Gong, W & Wang, JM. 1996. Human recombinant monocyte chemotactic protein and other C-C chemokines bind and induce directional migration of dendritic cells in vitro. *J Leukoc Biol* 60(3), pp. 365–71.

Ye, Q & Worman, HJ. 1994. Primary structure analysis and lamin B and DNA binding of human LBR, an integral protein of the nuclear envelope inner membrane. *J Biol Chem* 269(15), pp. 11306–11.

Young, MD, Wakefield, MJ, Smyth, GK & Oshlack, A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 11(2), p. R14.

Yu, B, Becnel, J, Zerfaoui, M, Rohatgi, R, Boulares, AH & Nichols, C. 2008. Serotonin 5-hydroxytryptamine 2A receptor activation suppresses tumor necrosis factor-$\alpha$-induced inflammation with extraordinary potency. *J Pharmacol Exp Ther* 327(2), pp. 316–23.

Zaidi, SK, Young, DW, Choi, JY, Pratap, J, Javed, A, Montecino, M, Stein, JL, Lian, JB, Van Wijnen, AJ & Stein, GS. 2004. Intranuclear trafficking: Organization and assembly of regulatory machinery for combinatorial biological control. *J Biol Chem* 279(42), pp. 43363–6.

Zaina, S, Pérez-Luque, EL & Lund, G. 2010. Genetics talks to epigenetics? The interplay between sequence variants and chromatin structure. *Curr Genomics* 11(5), pp. 359–67.

Zeisel, SH. 2007. Nutrigenomics and metabolomics will change clinical nutrition and public health practice: insights from studies on dietary requirements for choline. *Am J Clin Nutr* 86(3), pp. 542–8.

Zent, CS & Kay, NE. 2007. Chronic Lymphocytic Leukemia: Biology and Current Treatment. *Curr Oncol Rep* 9(5), pp. 345–52.

Zenz, T, Mertens, D, Küppers, R, Döhner, H & Stilgenbauer, S. 2010. From pathogenesis to treatment of chronic lymphocytic leukaemia. *Nat Rev Cancer* 10(1), pp. 37–50.

Zhang, W, Kater, AP, Widhopf, GF, Chuang, HY, Enzler, T, James, DF, Poustovoitov, M, Tseng, PH, Janz, S, Hoh, C, Herschman, H, Karin, M & Kipps, TJ. 2010. B-cell activating factor and v-Myc myelocytomatosis viral oncogene homolog (c-Myc) influence progression of chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A* 107(44), pp. 18956–60.

Zhu, J, Adli, M, Zou, JY, Verstappen, G, Coyne, M, Zhang, X, Durham, T, Miri, M, Deshpande, V, De Jager, PL, Bennett, DA, Houmard, JA, Muoio, DM, Onder, TT, Camahort, R, Cowan, CA, Meissner, A, Epstein, CB, Shoresh, N & Bernstein, BE. 2013. Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues. *Cell* 152(3), pp. 642–54.

# Annex I

## Supplementary Material

# Gene Expression Variability in CLL

Supplementary Figures

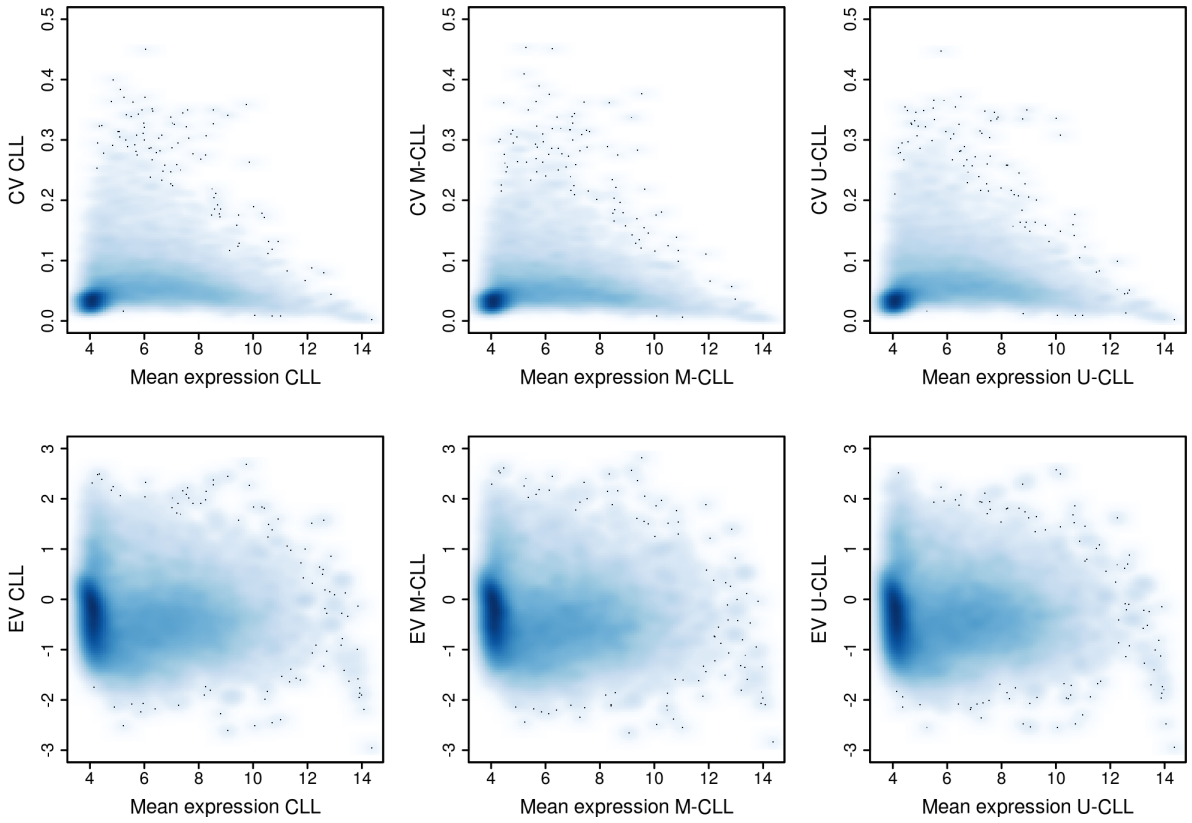


Figure SF1: Scatterplots of CV and EV distributions and their correlation with mean expression. Darker colors indicate an increased density of data points in the corresponding region of the plot. Top row: CV versus mean expression across all CLL samples, and across only M-CLL and U-CLL samples respectively. Bottom row: The same for the EV.

Figure taken from Ecker *et al.* (2015).

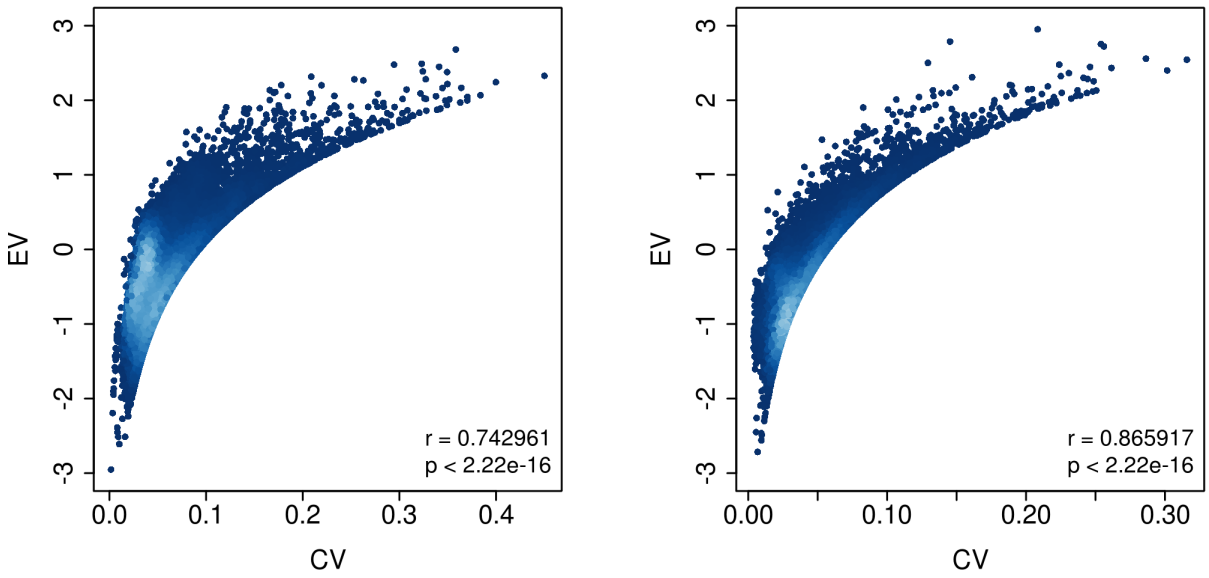


Figure SF2: Correlation of EV and CV in the data of the ICGC and Fabris. Lighter colors indicate higher densities of data points in the corresponding regions of the plot. Left panel: Scatterplot of EV versus CV using the ICGC dataset. Right panel: Scatterplot of EV versus CV using the Fabris dataset. Figure taken from Ecker *et al.* (2015).

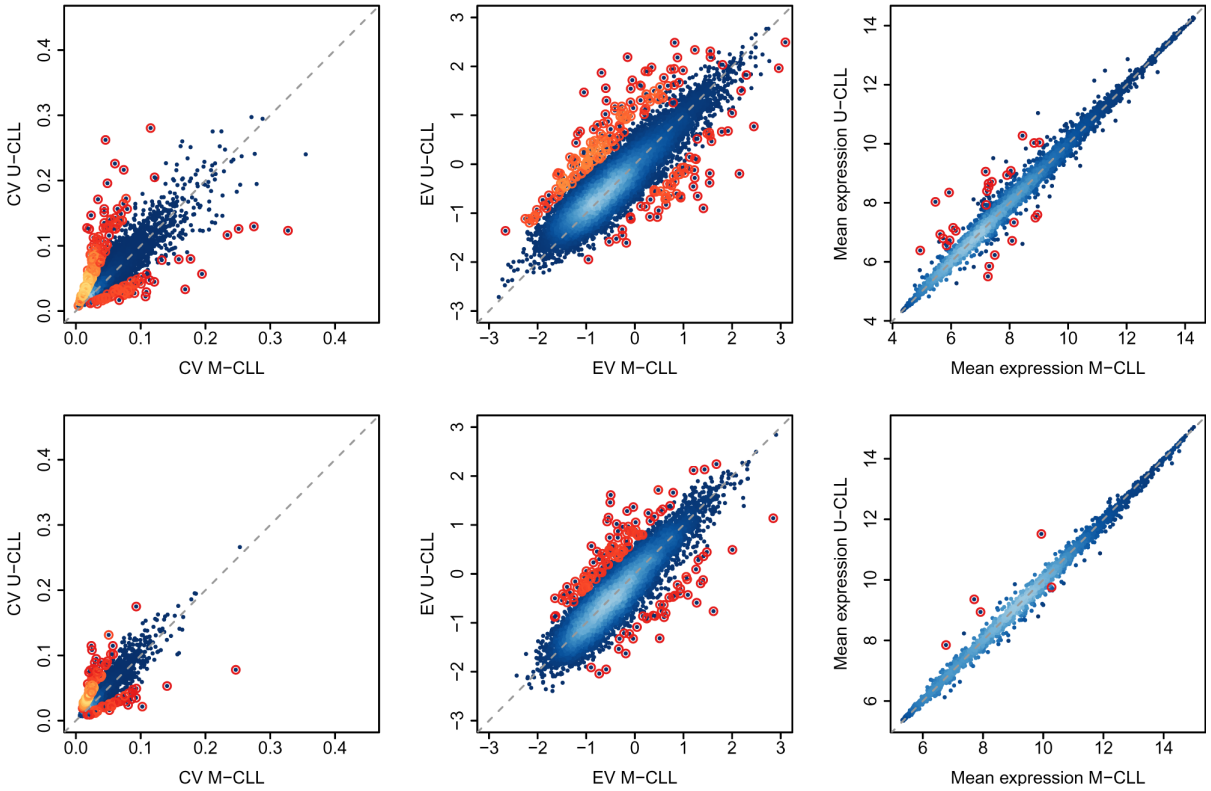


Figure SF3: Gene expression variability comparison of U-CLL and M-CLL in the datasets of Fabris and Haslinger. Lighter colors indicate higher densities of data points in the corresponding regions of the plot. Genes with statistically significant p-values at an FDR of 5% are highlighted. The gray dashed line represents the identity line. Left panel: Scatterplot of CV across patients in the two disease subtypes. Genes with statistically significant differential variability according to the F-test are highlighted. Middle panel: Scatterplot of EV across patients in the two disease subtypes. Genes with statistically significant differential variability according to the F-test are highlighted again. Right panel: Scatterplot of mean expression levels across patients in the two disease subtypes. Genes with statistically significant differential expression are highlighted. Top row: Data of Fabris. Bottom row: Data of Haslinger.
Figure taken from Ecker *et al.* (2015).

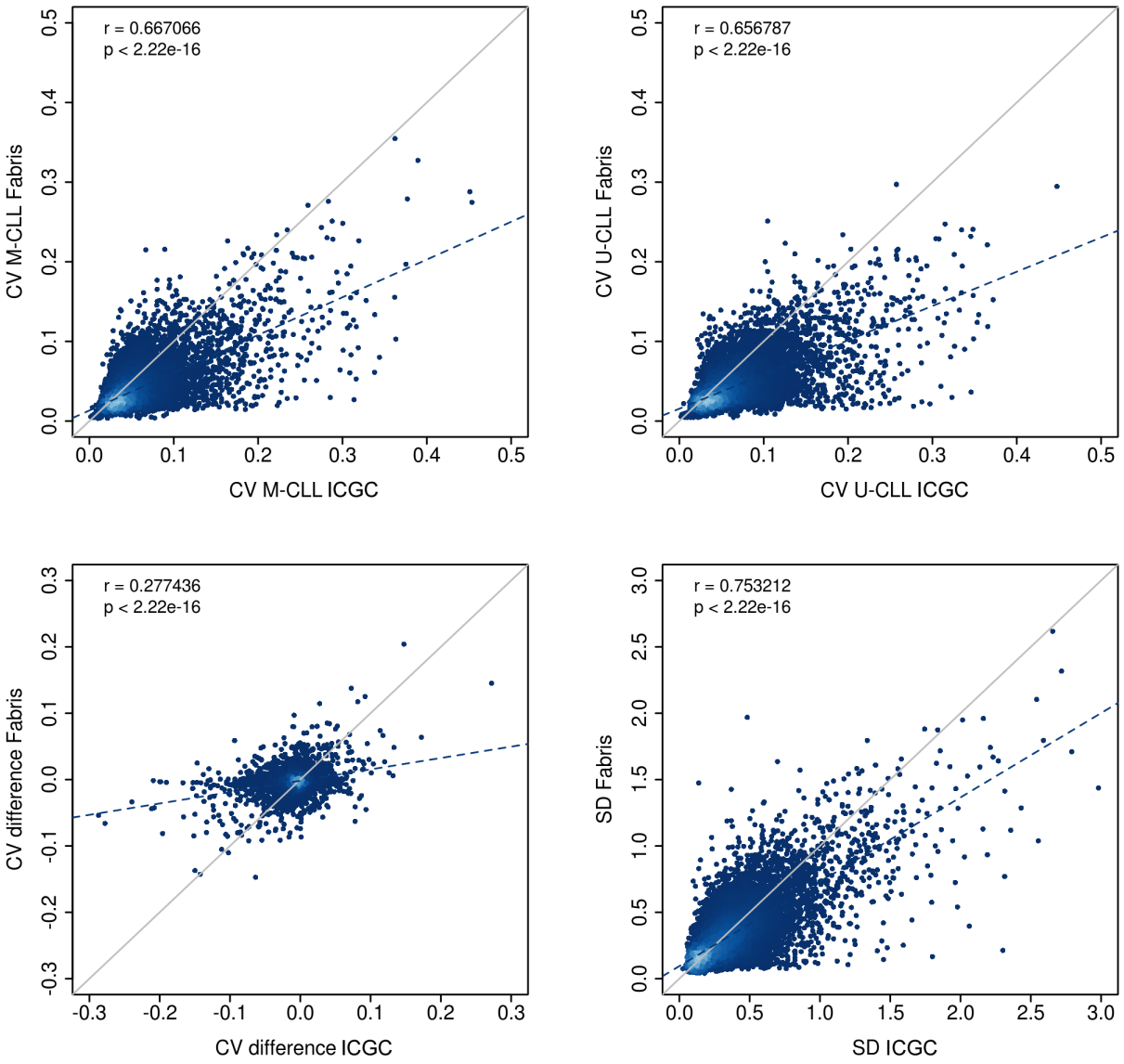


Figure SF4: Correlation of variability measurements between the ICGC data and the dataset of Fabris. Scatterplots comparing the two datasets. Lighter colors indicate higher densities of data points in the corresponding regions of the plot. The gray line represents the identity line, the blue dashed line represents the fitted regression line. Upper left panel: CV of M-CLL in Fabris versus CV of M-CLL in the ICGC data. Upper right panel: CV of U-CLL in Fabris versus CV of U-CLL in the ICGC data. Lower left panel: CV difference ($CV_{M\text{-}CLL} - CV_{U\text{-}CLL}$) in Fabris versus CV difference in the ICGC data. Lower right panel: SD across all CLL samples in Fabris versus SD in the ICGC data.
Figure taken from Ecker *et al.* (2015).

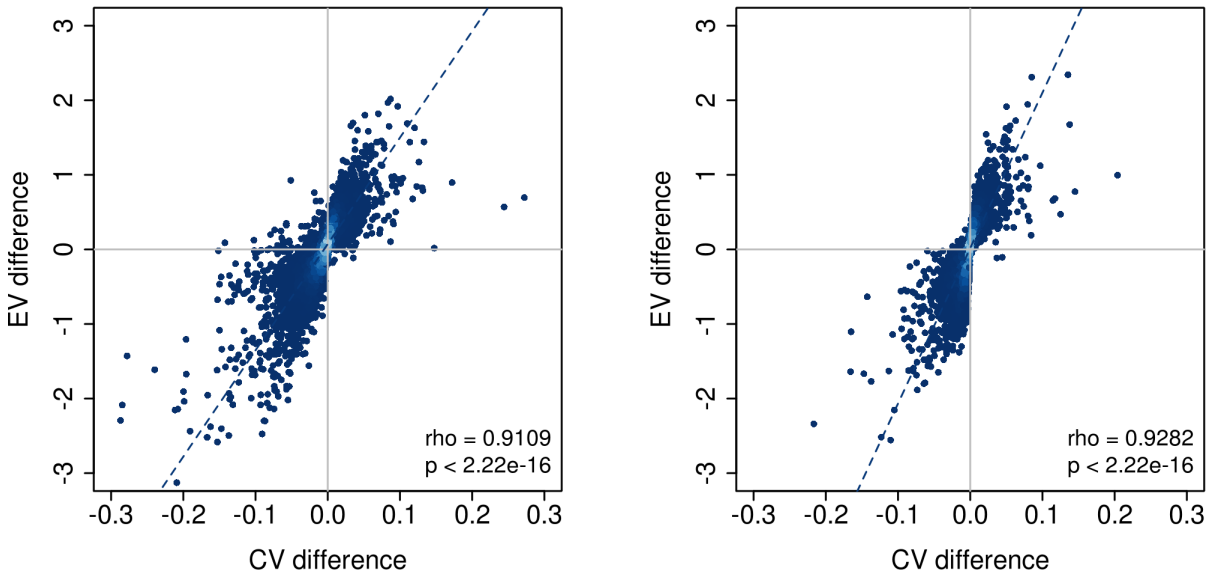


Figure SF5: Correlation of EV difference and CV difference in the ICGC data and the dataset of Fabris. EV difference = $EV_{M\text{-}CLL} - EV_{U\text{-}CLL}$ and CV difference = $CV_{M\text{-}CLL} - CV_{U\text{-}CLL}$. Lighter colors indicate higher densities of data points in the corresponding regions of the plot. The blue dashed line represents the fitted regression line. Left panel: Scatterplot of EV difference versus CV difference using the ICGC dataset. Right panel: Scatterplot of EV difference versus CV difference using the Fabris data.
Figure taken from Ecker *et al.* (2015).

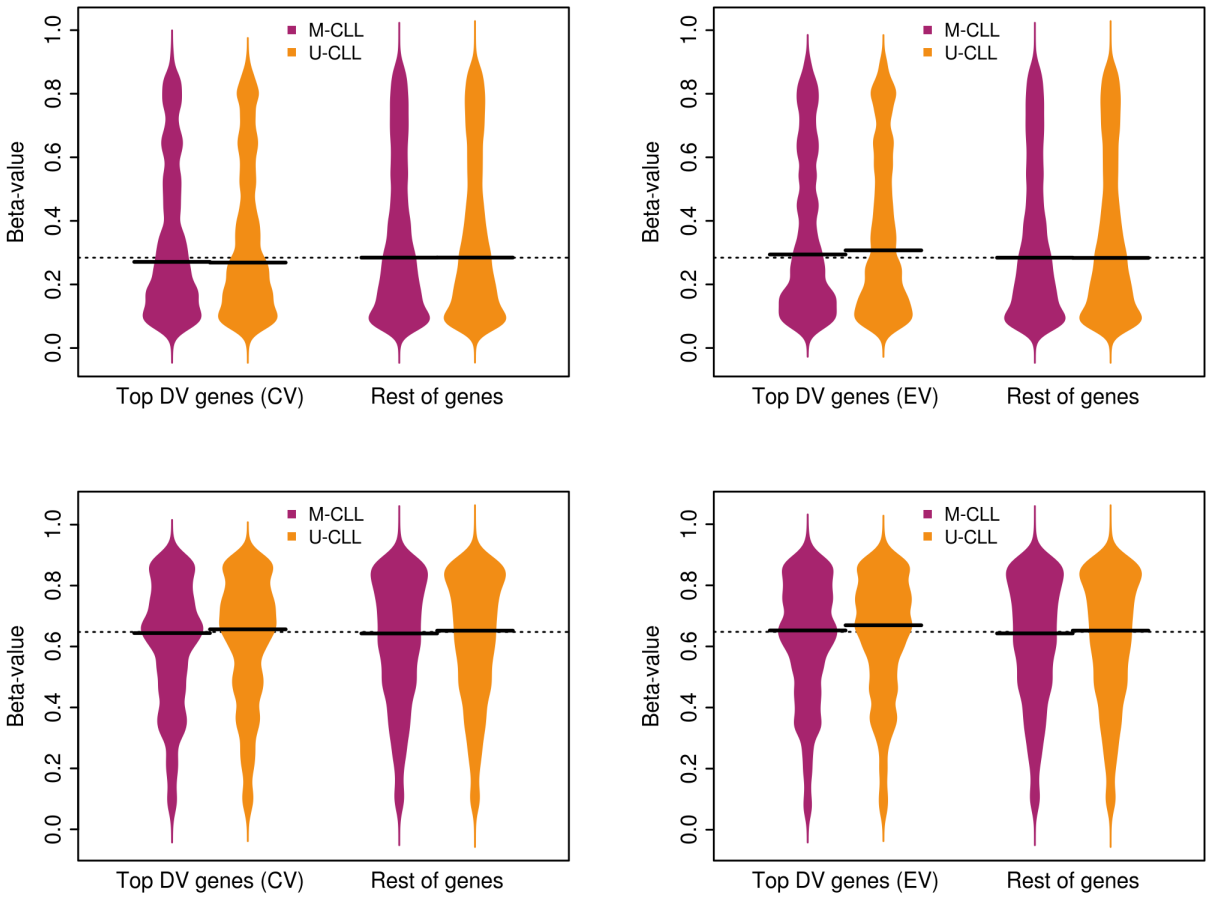


Figure SF6: Beanplots comparing the methylation profiles of M-CLL and U-CLL of the top 500 genes with increased variability in U-CLL. Methylation measurements are given in beta values. Top row: Promoter methylation. Bottom row: Gene body methylation. Left panel: Differential variability based on CV differences. Right panel: Differential variability based on EV differences.
Figure taken from Ecker *et al.* (2015).

# Variability in Normal Blood Cells

Supplementary Figures

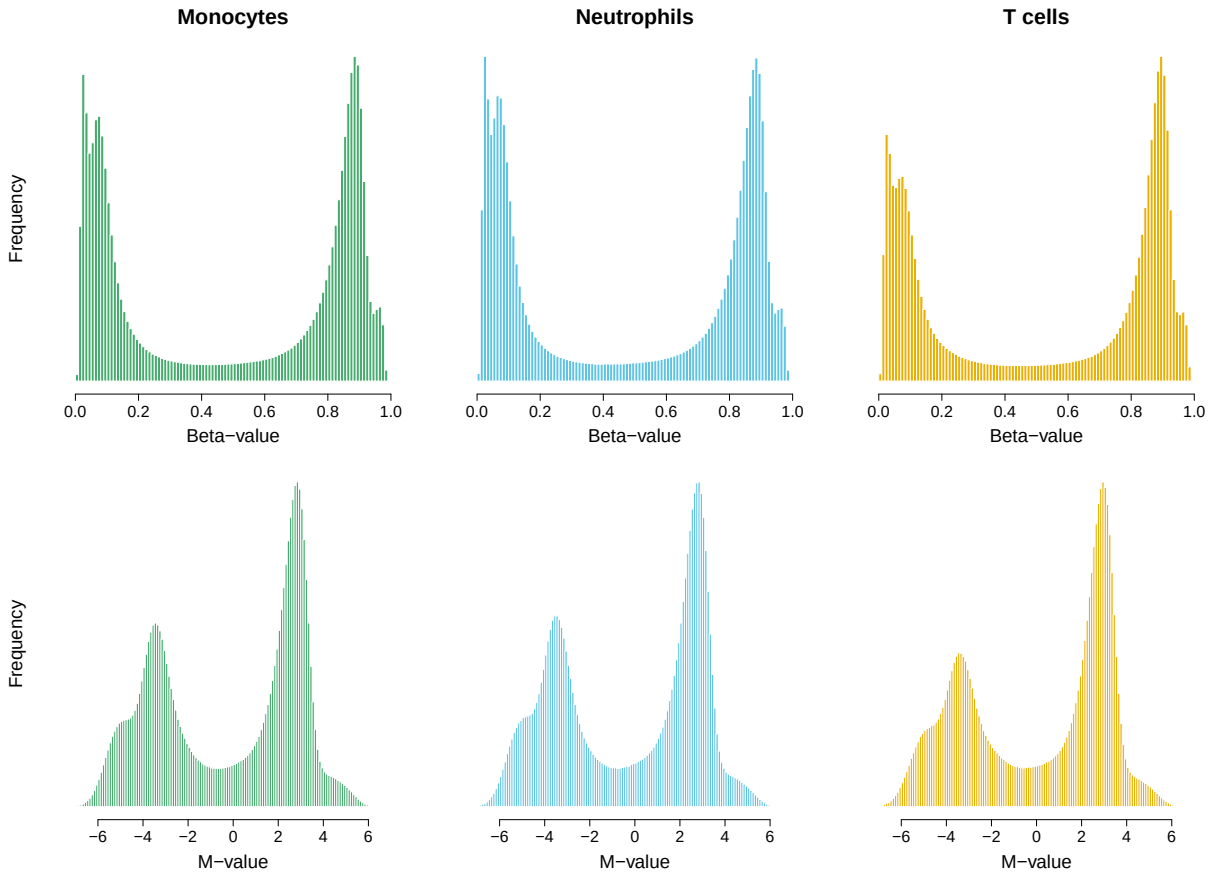


Figure SF7: Distribution of Beta-values and M-values in the three cell types. Top row: Beta-value distribution in monocytes, neutrophils and T cells. Bottom row: M-value distribution in monocytes, neutrophils and T cells.

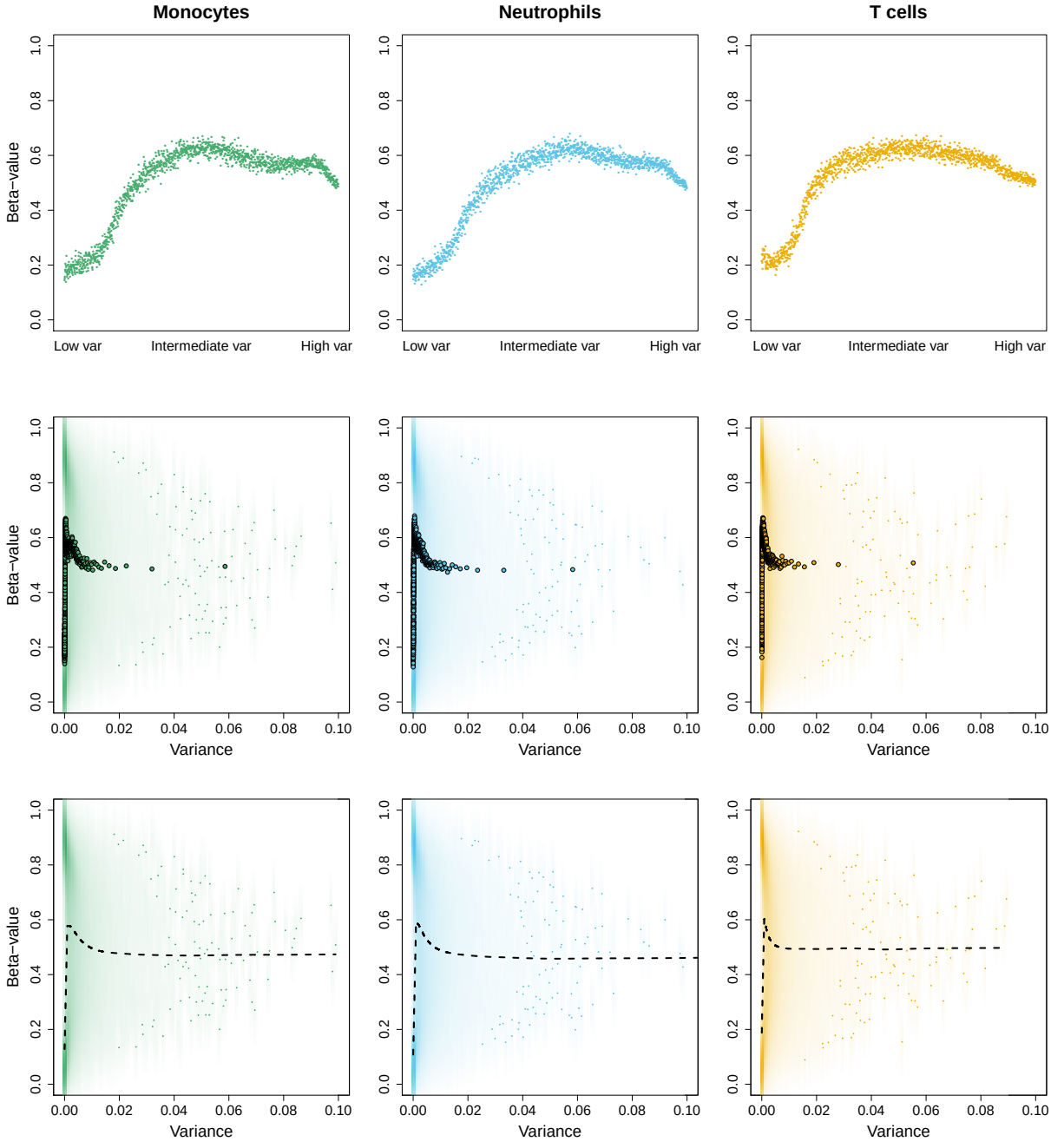


Figure SF8: Mean Beta-values versus variance of Beta-values. Top row: CpG-wise variances of Beta-values were calculated. Then the values were ordered from low to high variance, grouped together in bins of 300 CpGs, and plotted against the mean Beta-value maintaining the ordering by variance, to see if variance is evenly distributed across Beta-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line. When calculating median Beta-values instead of mean Beta-values the observed patterns are almost perfectly the same (not shown).

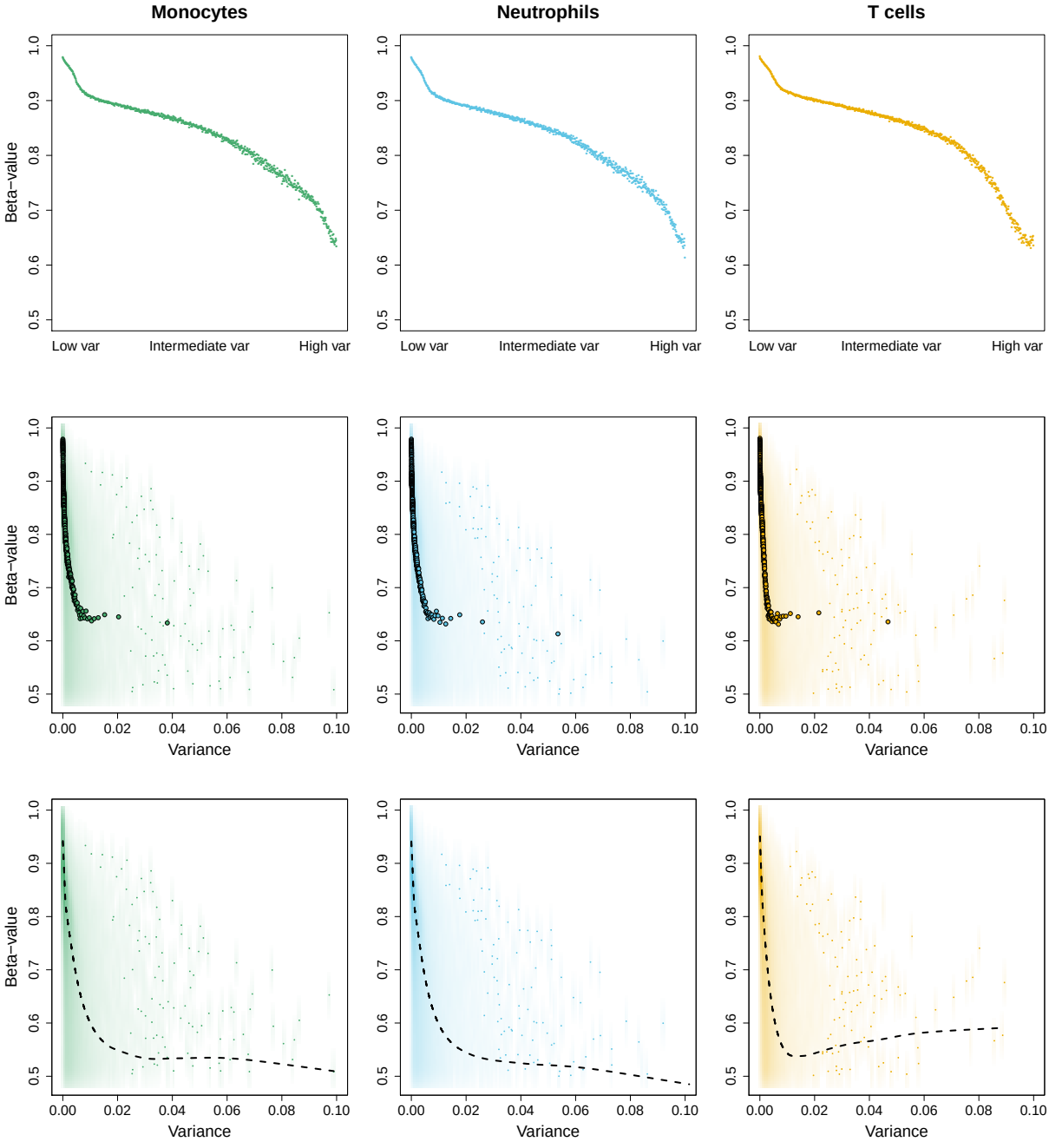


Figure SF9: Mean Beta-values versus variance of Beta-values in high Beta-values ($\geq 0.5$). Top row: CpG-wise variances of Beta-values were calculated. Then the values were ordered from low to high variance, grouped together in bins of 300 CpGs, and plotted against the mean Beta-value maintaining the ordering by variance, to see if variance is evenly distributed across Beta-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line. When calculating median Beta-values instead of mean Beta-values the observed patterns are almost perfectly the same (not shown).

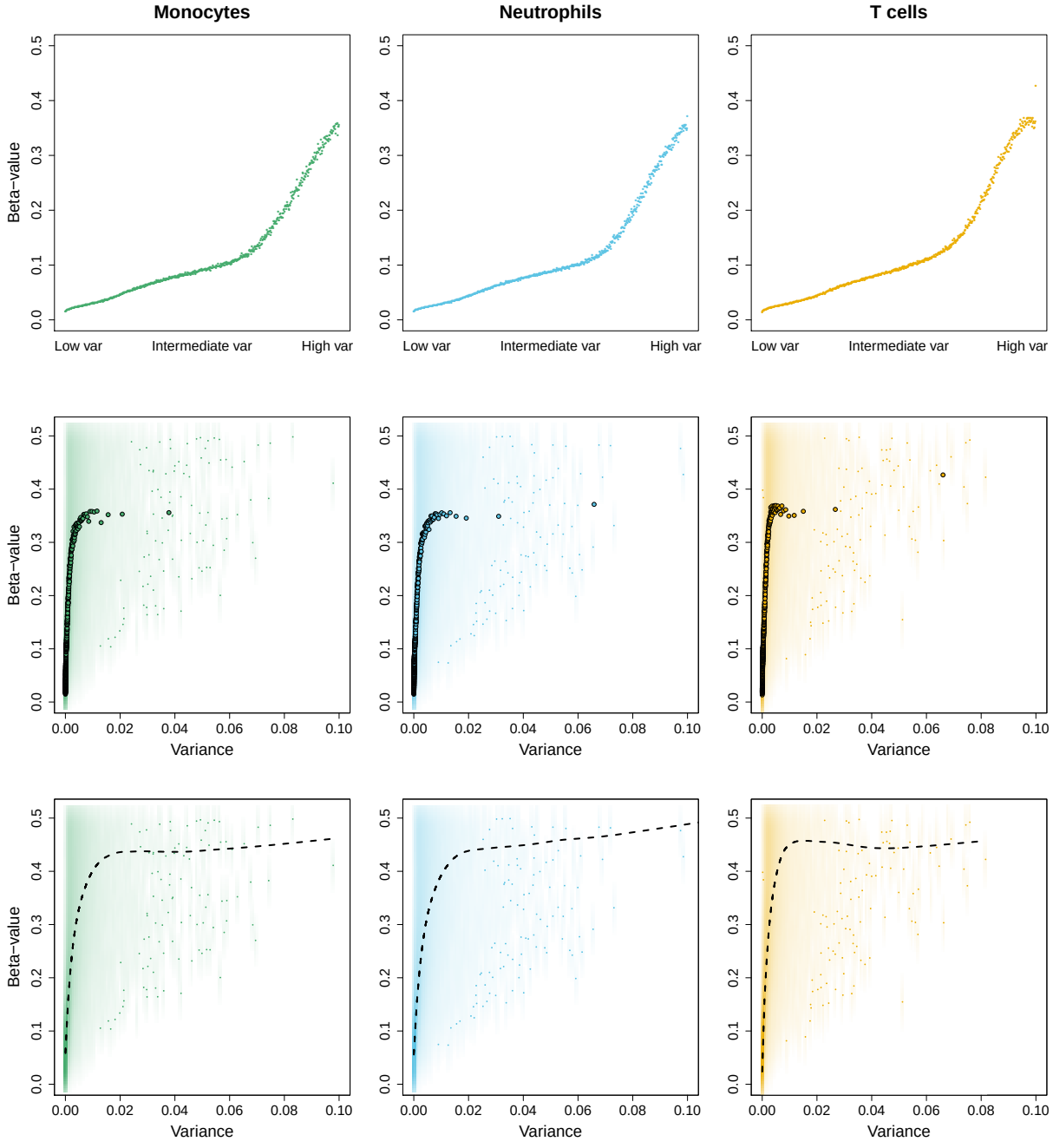


Figure SF10: Mean Beta-values versus variance of Beta-values in low Beta-values ($< 0.5$). Top row: CpG-wise variances of Beta-values were calculated. Then the values were ordered from low to high variance, grouped together in bins of 300 CpGs, and plotted against the mean Beta-value maintaining the ordering by variance, to see if variance is evenly distributed across Beta-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line. When calculating median Beta-values instead of mean Beta-values the observed patterns are almost perfectly the same (not shown).

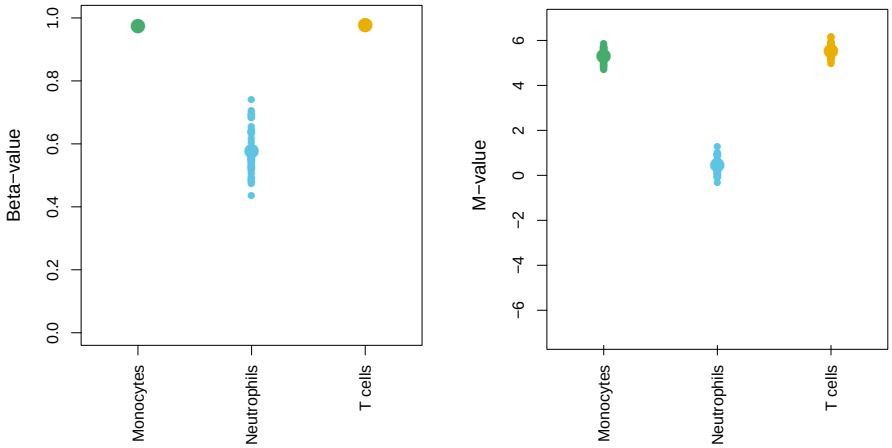


Figure SF11: Beta-values and M-values of probe cg07804973. Every data point represents the methylation value of one individual. The bigger data points correspond to the mean methylation value. Methylation values of this probe in neutrophils seem to be more spread out and therefore more variable in neutrophils when looking at Beta-values, but not when using M-values, which indicate that the locus is less methylated in neutrophils, but not showing much higher methylation variability than monocytes or T cells.

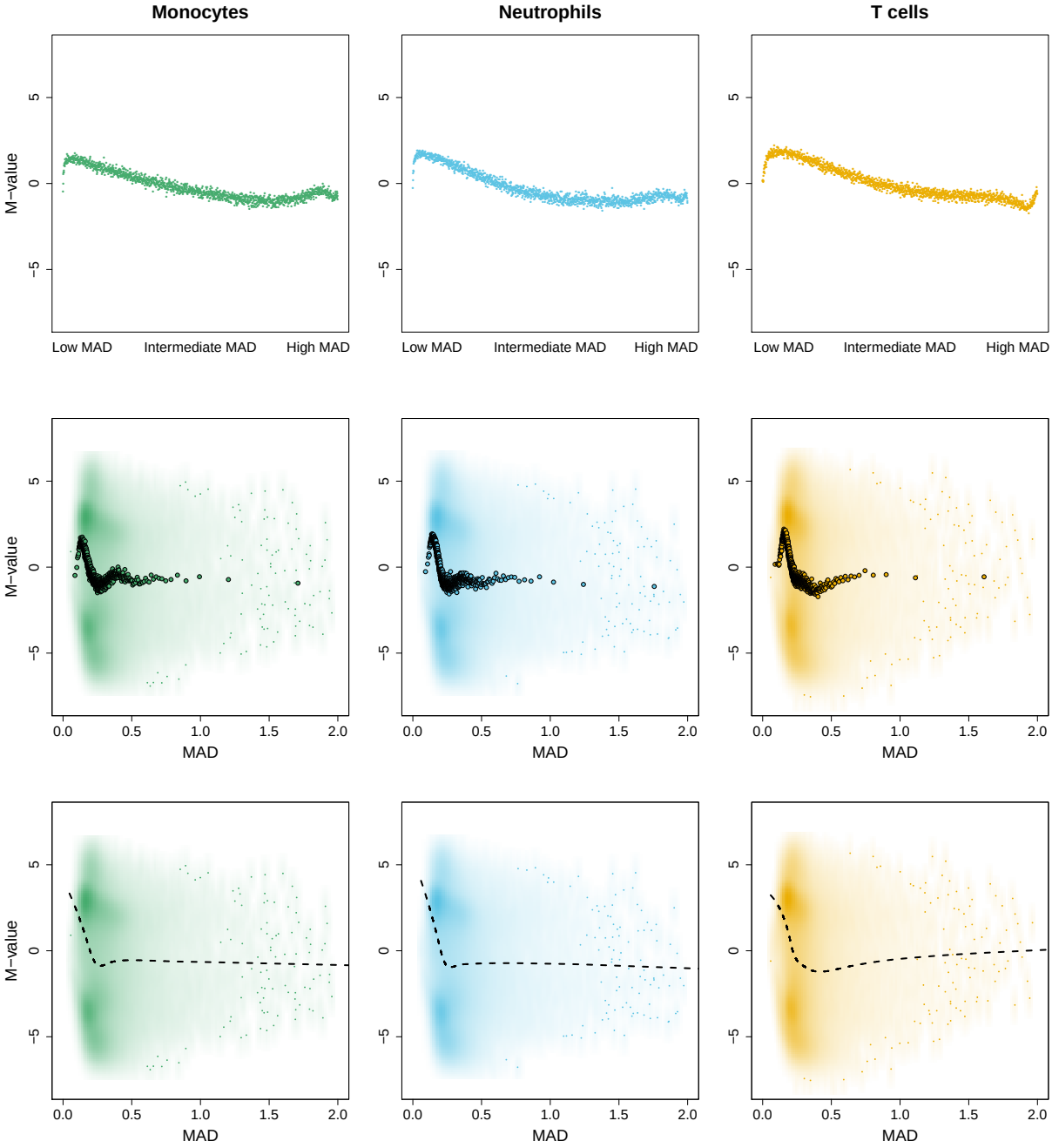


Figure SF12: Mean M-values versus MAD of M-values. Top row: CpG-wise MAD-values of M-values were calculated. Then the values were ordered from low to high MAD, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by MAD, to see if variability in terms of MAD-values is evenly distributed across M-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

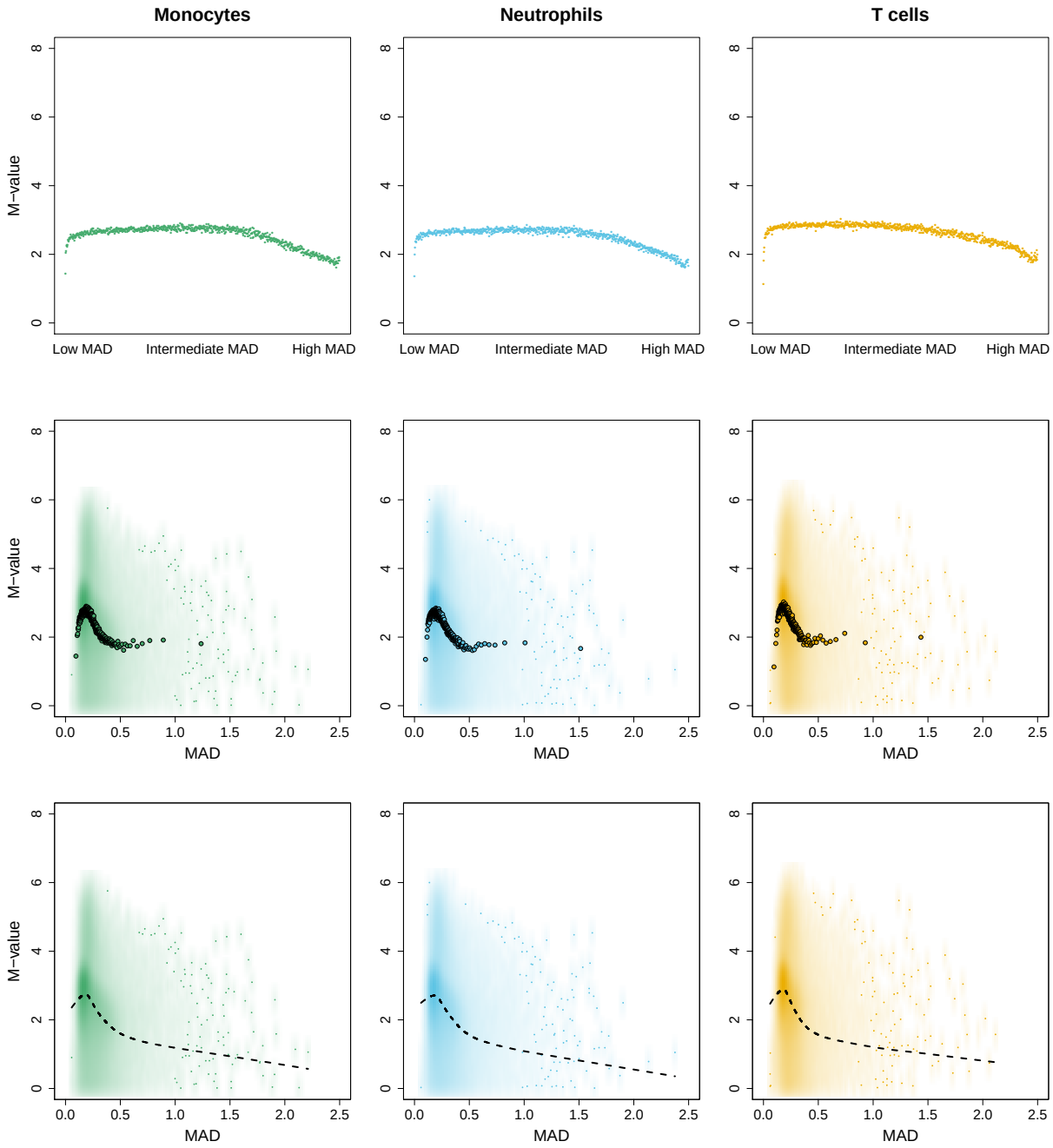


Figure SF13: Mean M-values versus MAD of M-values in positive M-values ($\geq 0$). Top row: CpG-wise MAD-values of M-values were calculated. Then the values were ordered from low to high MAD, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by MAD, to see if variability in terms of MAD-values is evenly distributed across M-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

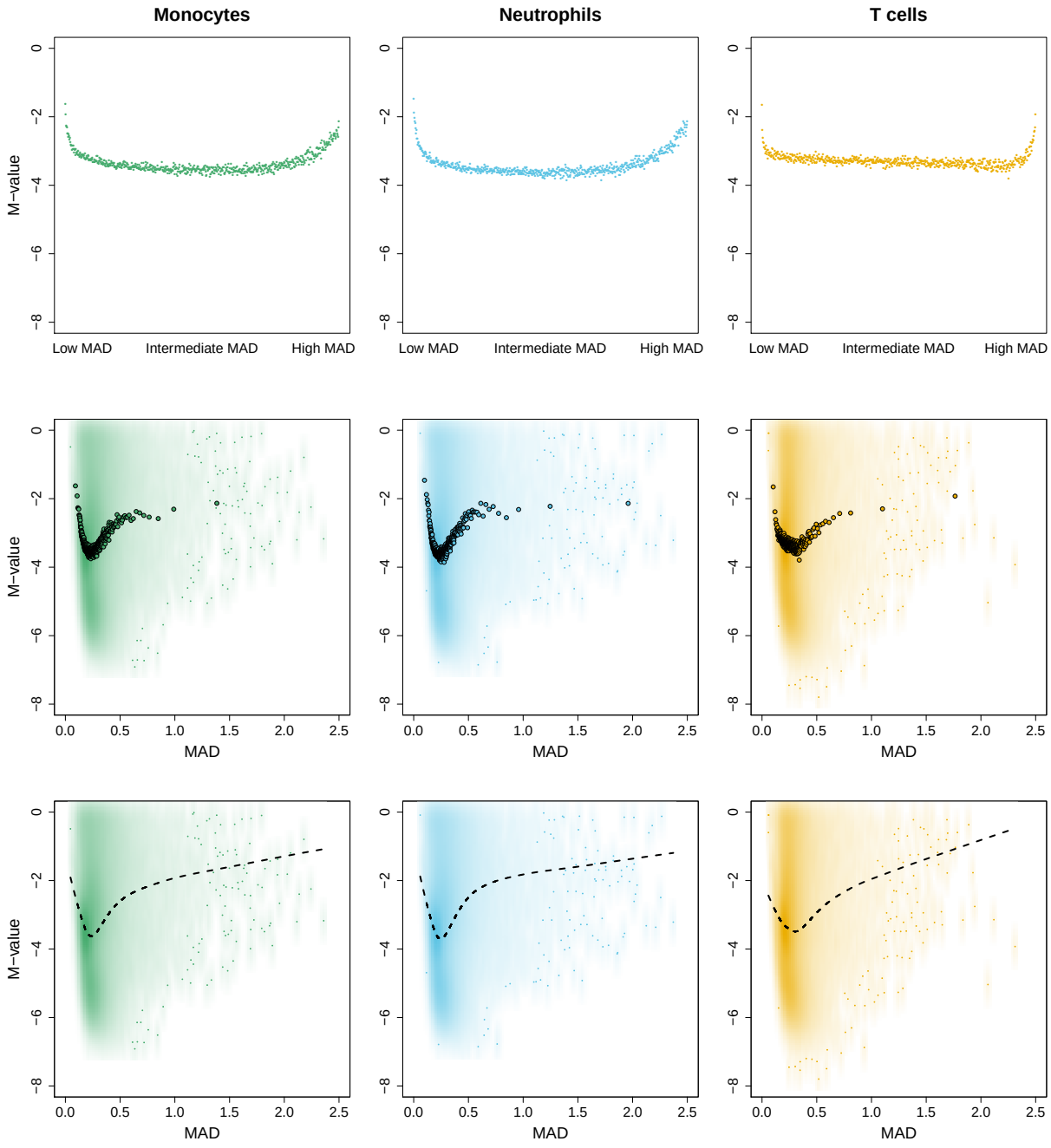


Figure SF14: Mean M-values versus MAD of M-values in negative M-values ($< 0$). Top row: CpG-wise MAD-values of M-values were calculated. Then the values were ordered from low to high MAD, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by MAD, to see if variability in terms of MAD-values is evenly distributed across M-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

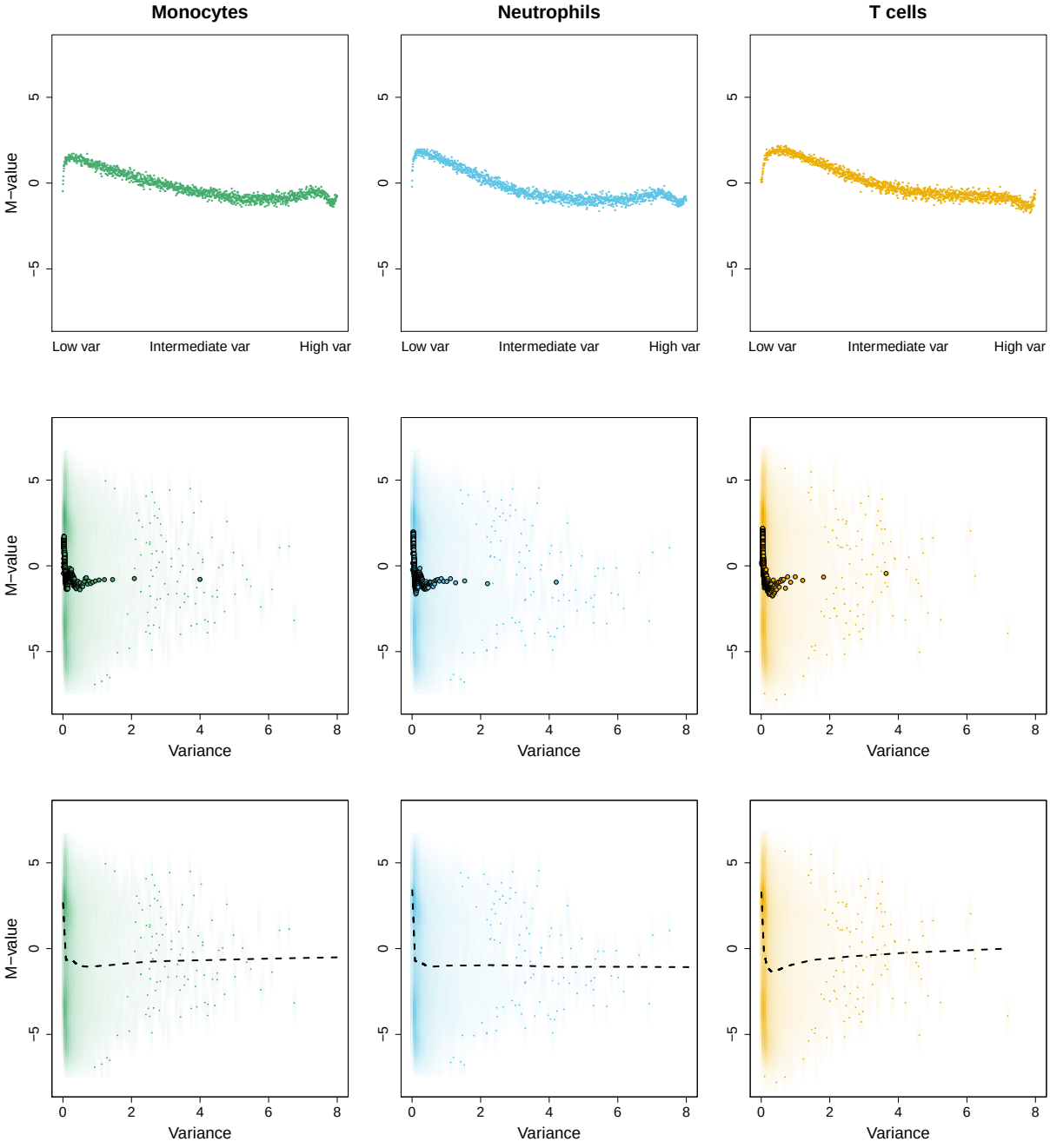


Figure SF15: Mean M-values versus variance of M-values. Top row: CpG-wise variances of M-values were calculated. Then the values were ordered from low to high variance, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by variance, to see if variance is evenly distributed across M-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

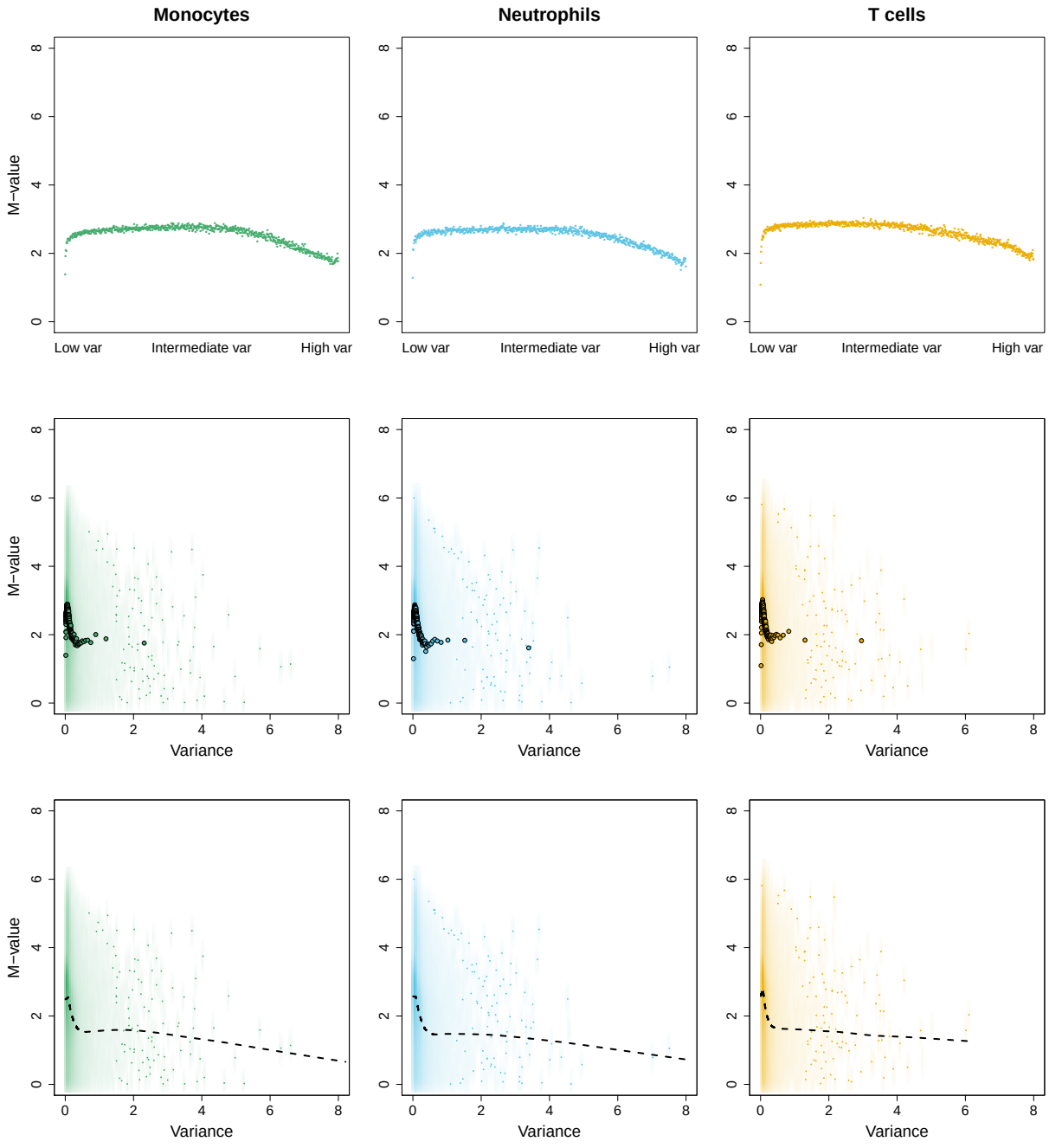


Figure SF16: Mean M-values versus variance of M-values in positive M-values ($\geq 0$). Top row: CpG-wise variances of M-values were calculated. Then the values were ordered from low to high variance, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by variance, to see if variance is evenly distributed across M-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

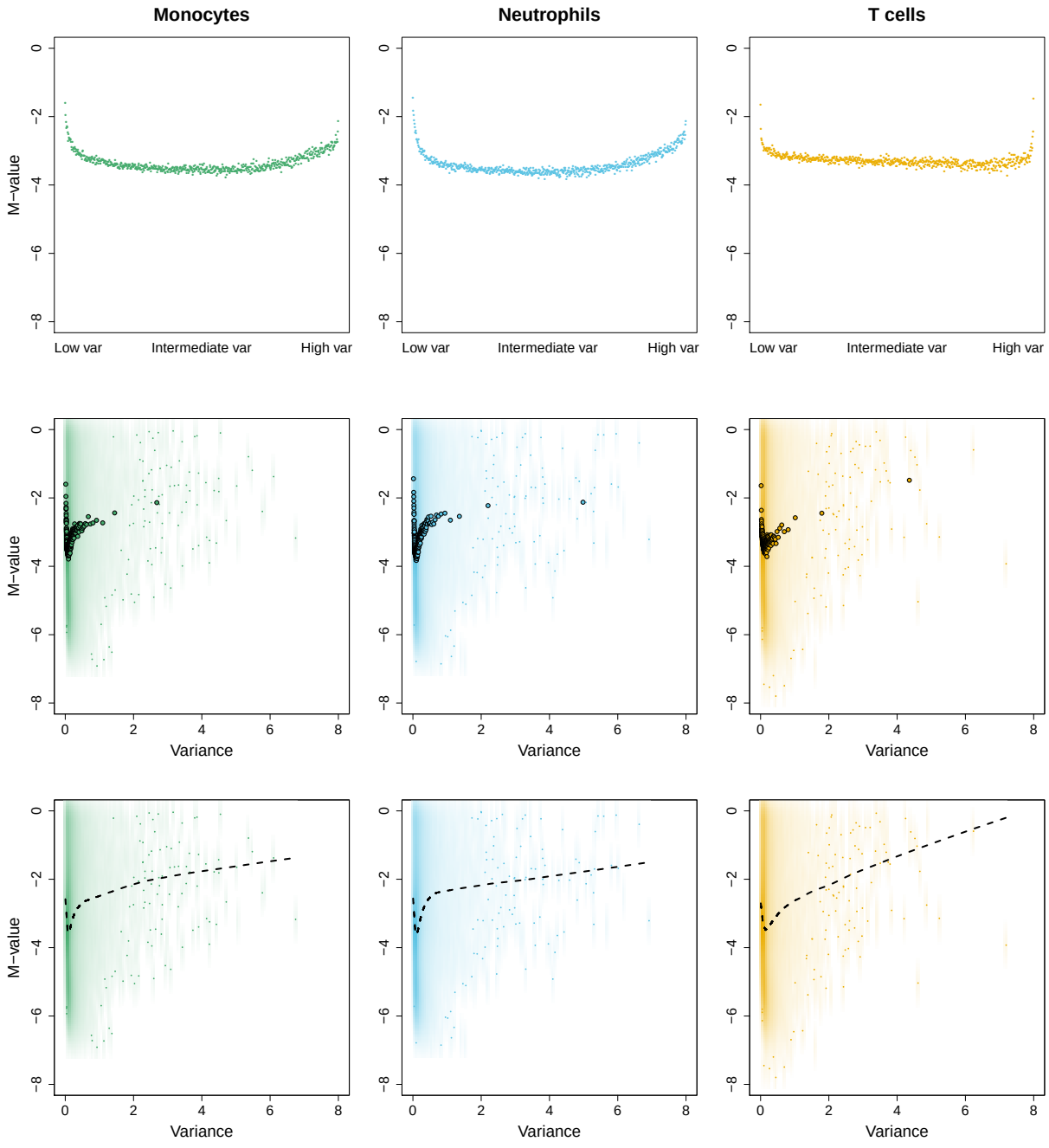


Figure SF17: Mean M-values versus variance of M-values in negative M-values ($< 0$). Top row: CpG-wise variances of M-values were calculated. Then the values were ordered from low to high variance, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by variance, to see if variance is evenly distributed across M-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

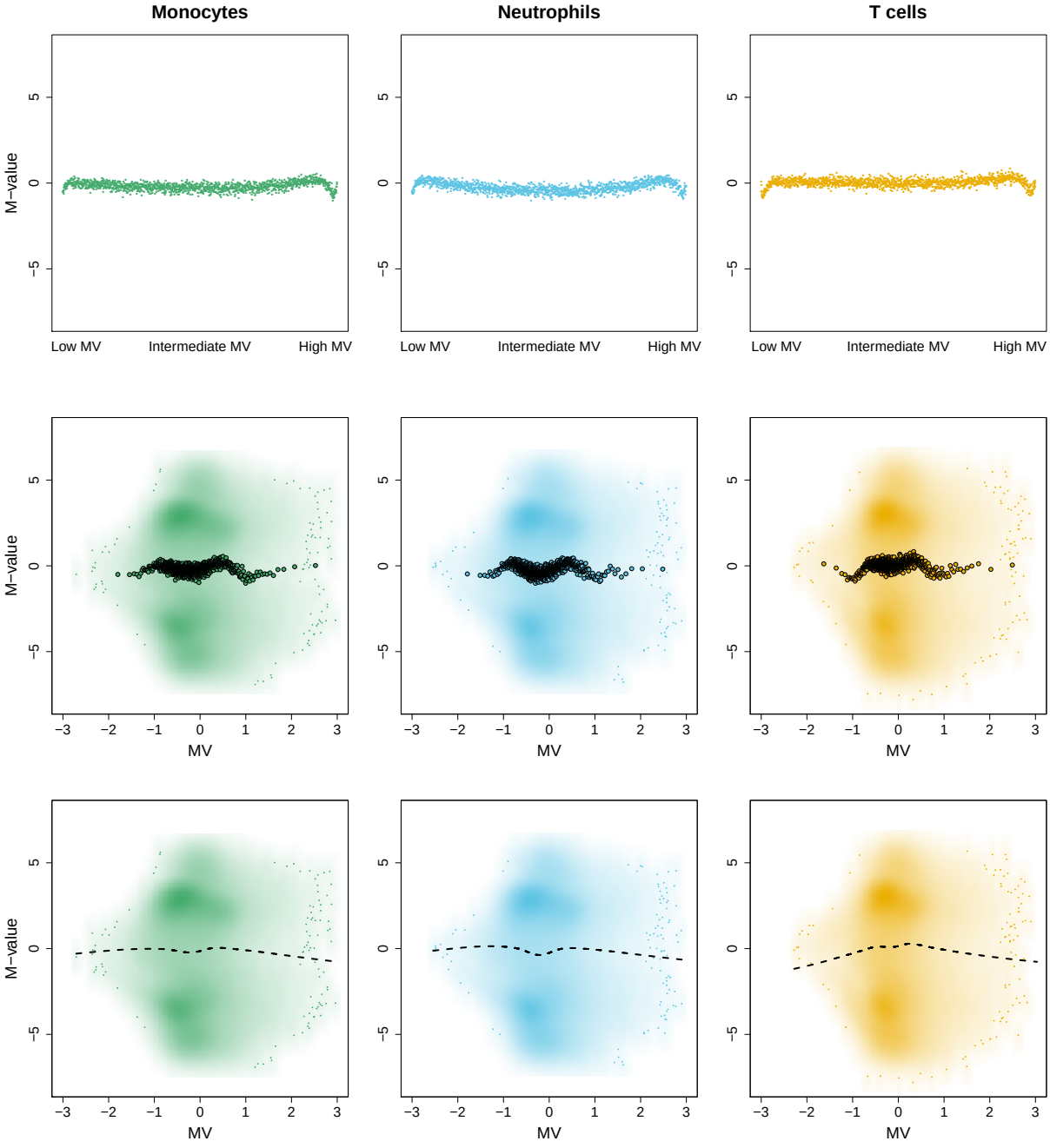


Figure SF18: Mean M-values versus MV. Top row: CpG-wise MV-scores were calculated. Then the values were ordered from low to high MV, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by MV, to see if the MV-score is evenly distributed across M-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

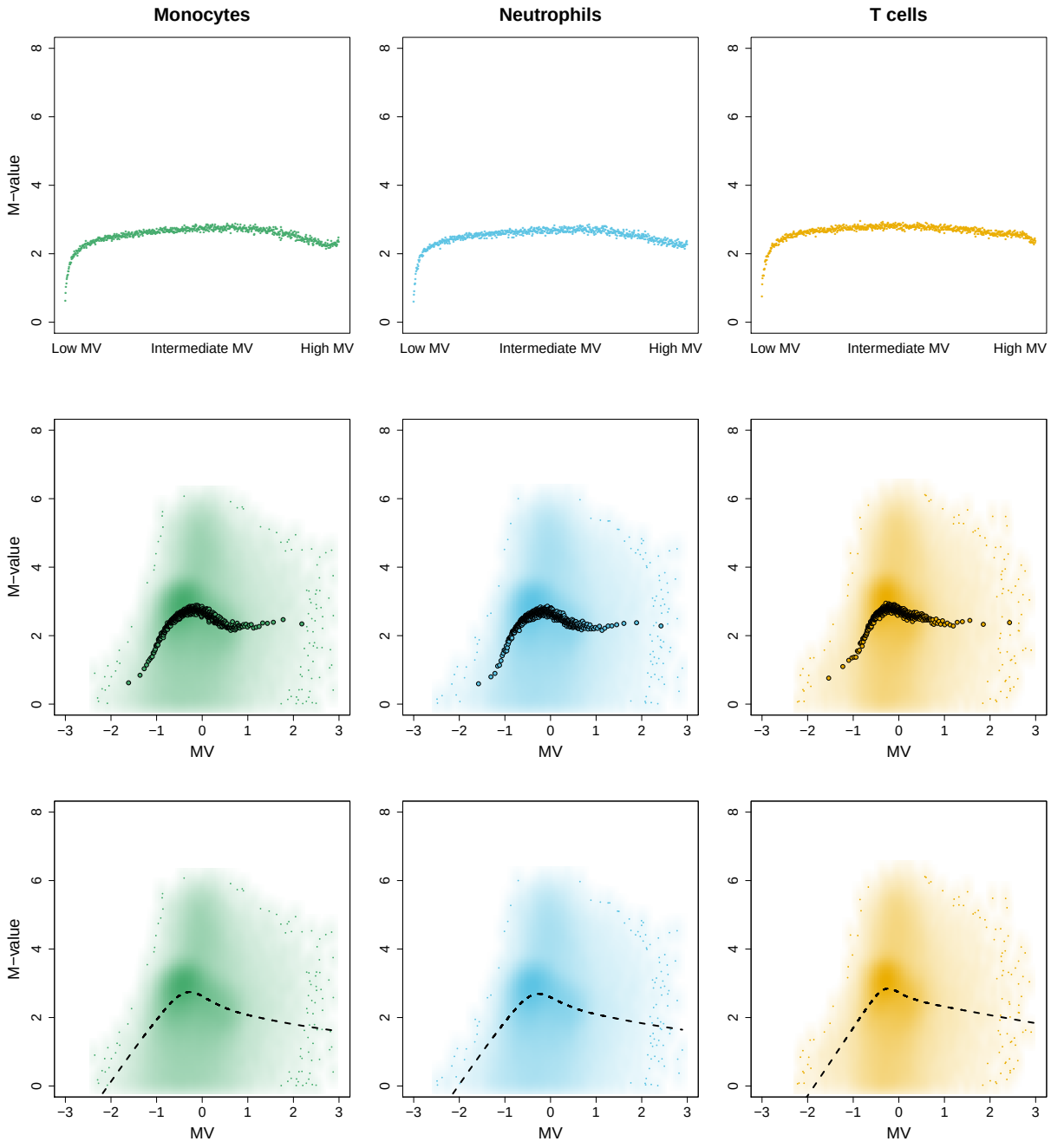


Figure SF19: Mean M-values versus MV in positive M-values ($\geq 0$). Top row: CpG-wise MV-scores were calculated. Then the values were ordered from low to high MV, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by MV, to see if the MV-score is evenly distributed across M-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

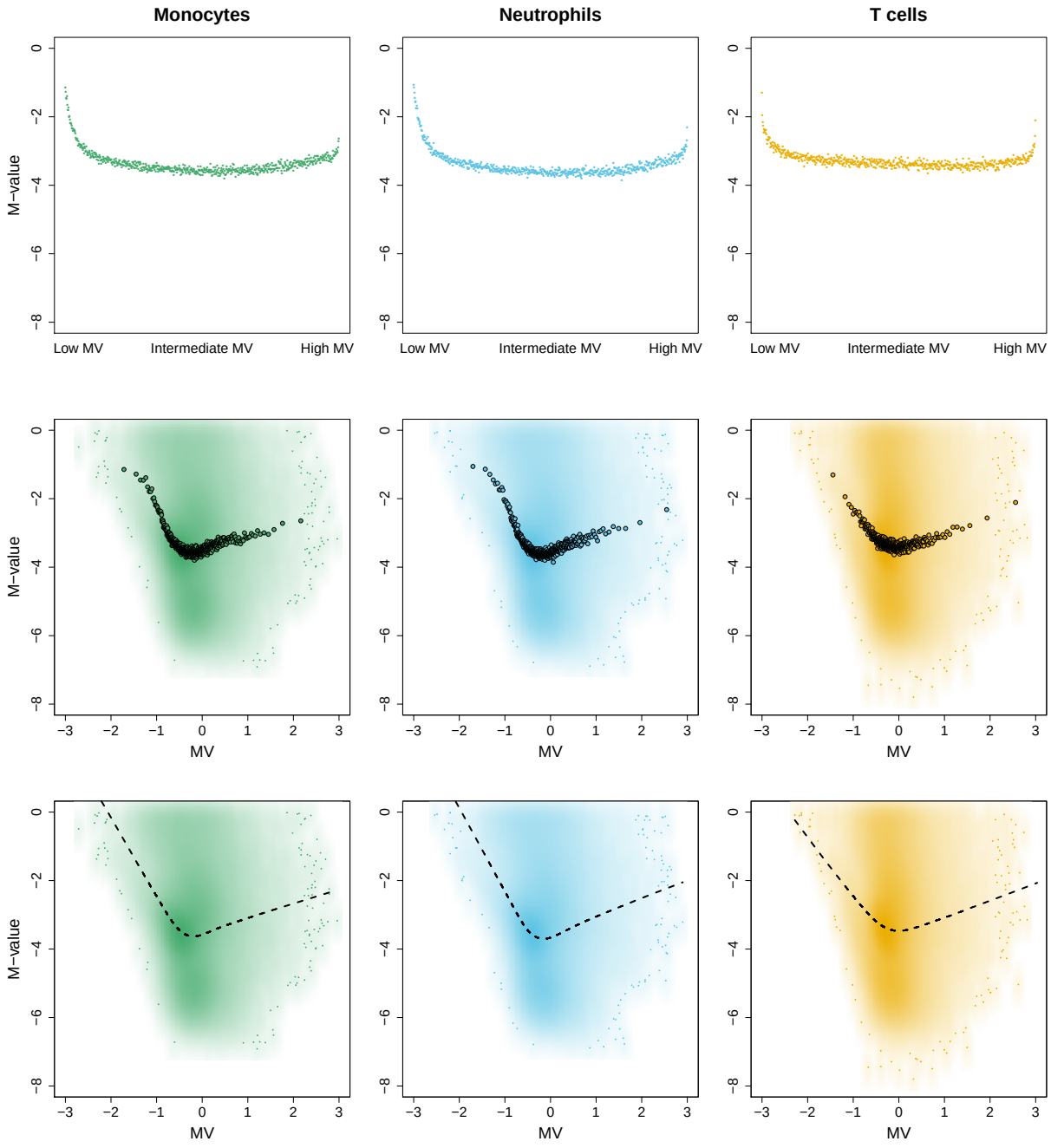


Figure SF20: Mean M-values versus MV in negative M-values ($< 0$). Top row: CpG-wise MV-scores were calculated. Then the values were ordered from low to high MV, grouped together in bins of 300 CpGs, and plotted against the mean M-value maintaining the ordering by MV, to see if the MV-score is evenly distributed across M-values. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

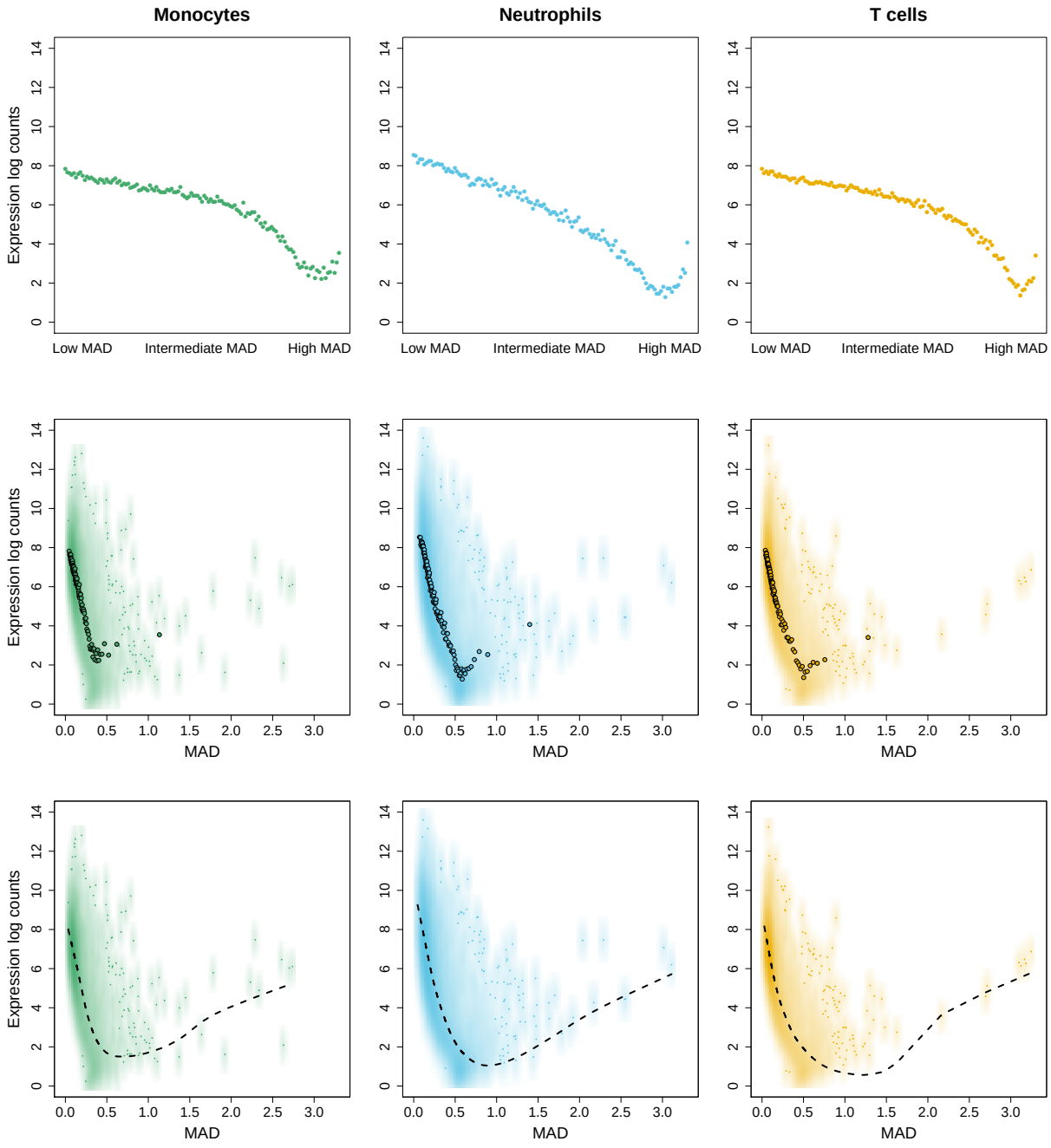


Figure SF21: Mean expression values versus MAD. Top row: Gene-wise MAD-values were calculated. Then the values were ordered from low to high MAD, grouped together in bins of 100 genes, and plotted against mean expression log counts maintaining the ordering by MAD-values, to see if the MAD is evenly distributed across expression levels. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

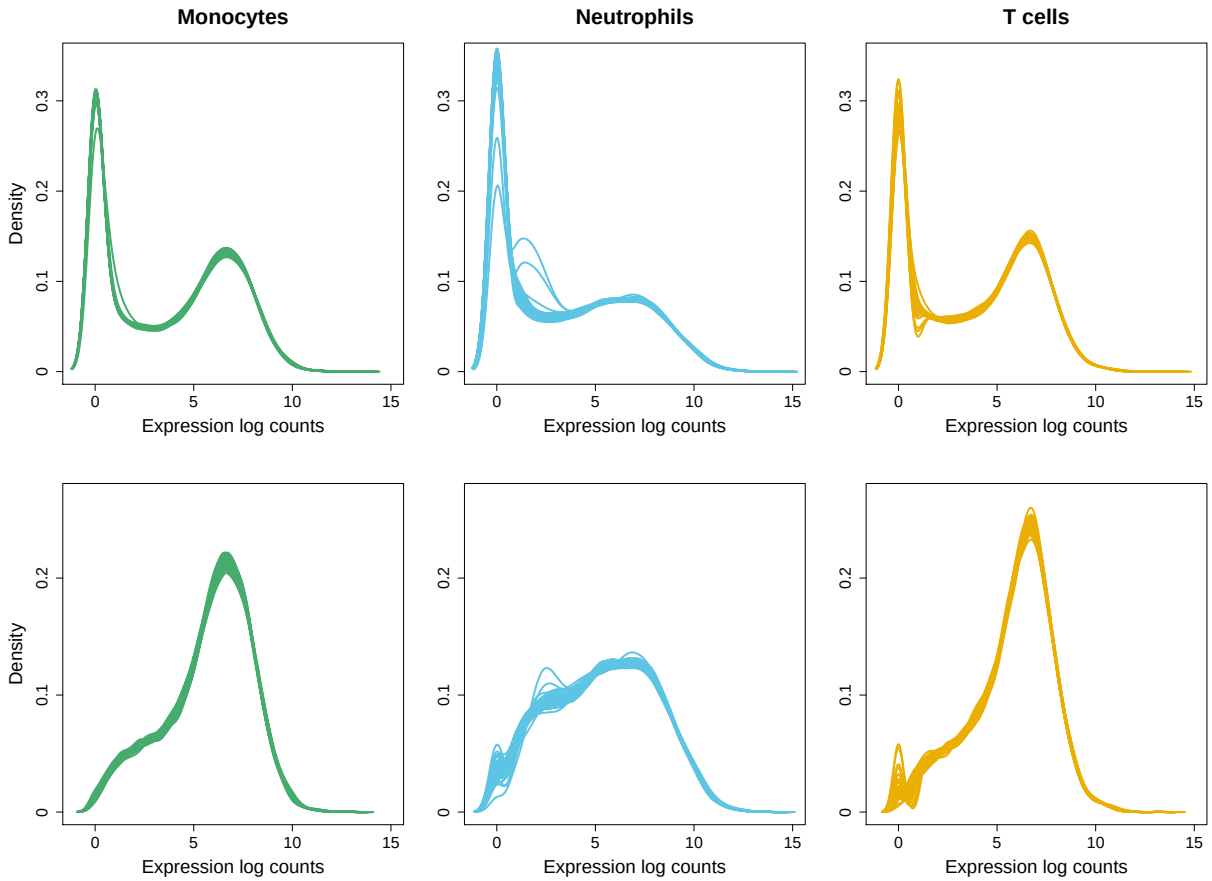


Figure SF22: Distribution of expression values in the three cell types. Top row: Normalized gene expression values in log counts, only protein coding genes. Bottom row: The same as above, after the filtering applied for the analysis of differential variability, removing all genes with no reads in more than 50% of the samples in one or more of the groups (see section 3.2.5).

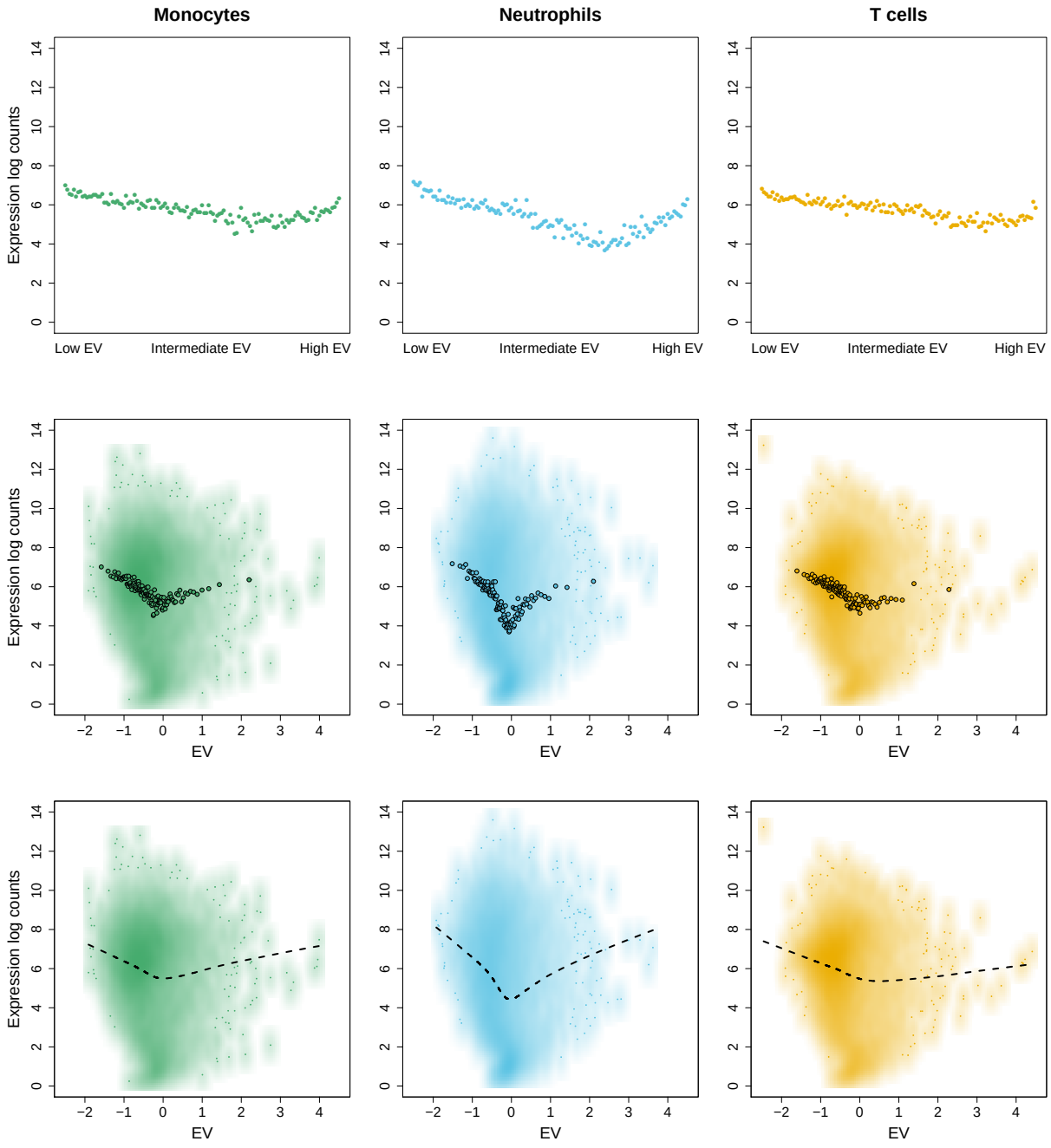


Figure SF23: Mean expression values versus EV. Top row: Gene-wise EV-values were calculated. Then the values were ordered from low to high EV, grouped together in bins of 100 genes, and plotted against mean expression log counts maintaining the ordering by EV-values, to see if the EV is evenly distributed across expression levels. Middle row: Scatterplots of the original data values, including the binned data points from the plots in the first row (filled circles). Bottom row: Scatterplots of the original data values, with a lowess regression line.

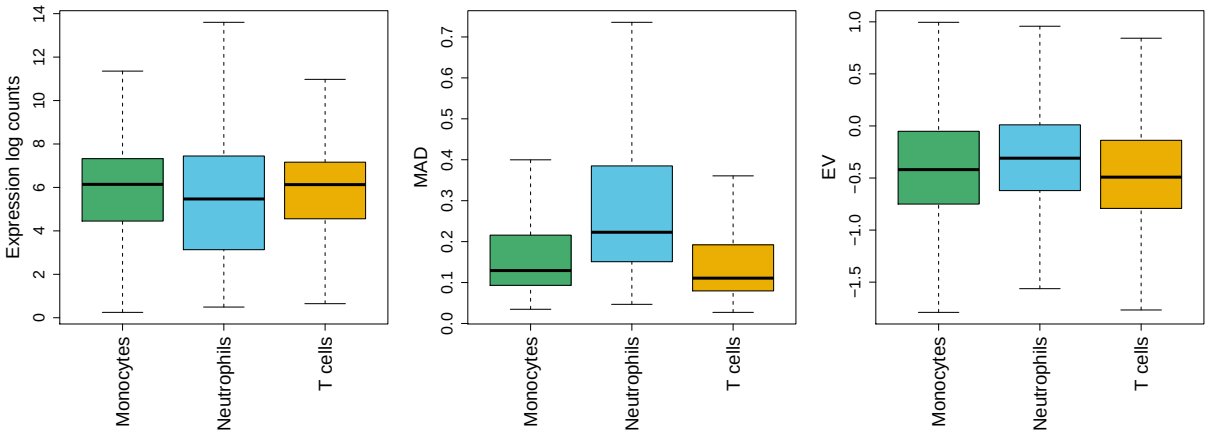


Figure SF24: Distribution of mean expression and expression variability measurements in the three cell types. Left: Mean expression levels (see also supplementary figure SF22). Middle: MAD-values of expression. Right: EV-values of expression.

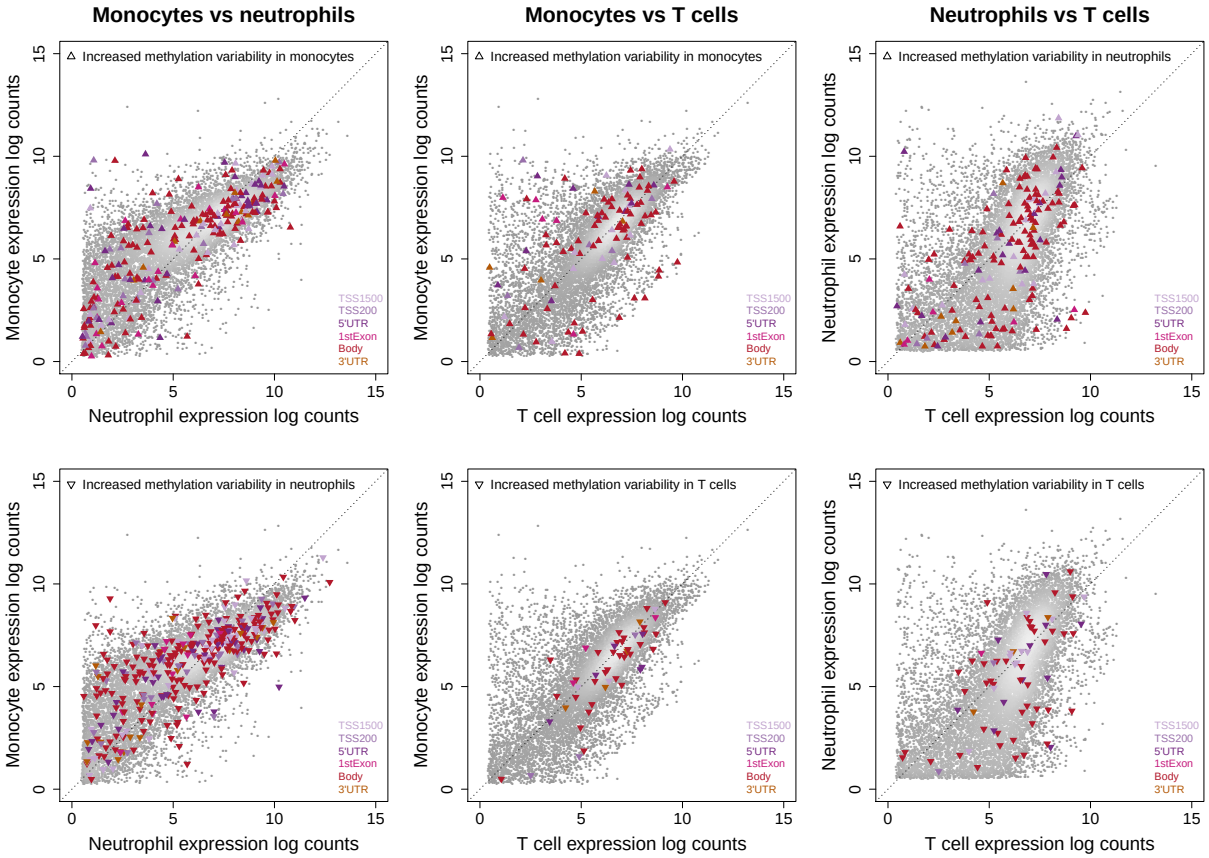


Figure SF25: Scatterplots of mean expression levels marking genes with differential DNA methylation variability. Every data point represents a gene. The colors correspond to the genomic region to which the hypervariably methylated CpGs of the genes belong. The dotted line represents the identity line. The three different pair-wise comparisons of the three groups are shown in columns. Top row: Genes with increased DNA methylation variability in the first group are marked. Bottom row: Genes with increased DNA methylation variability in the second group are marked.

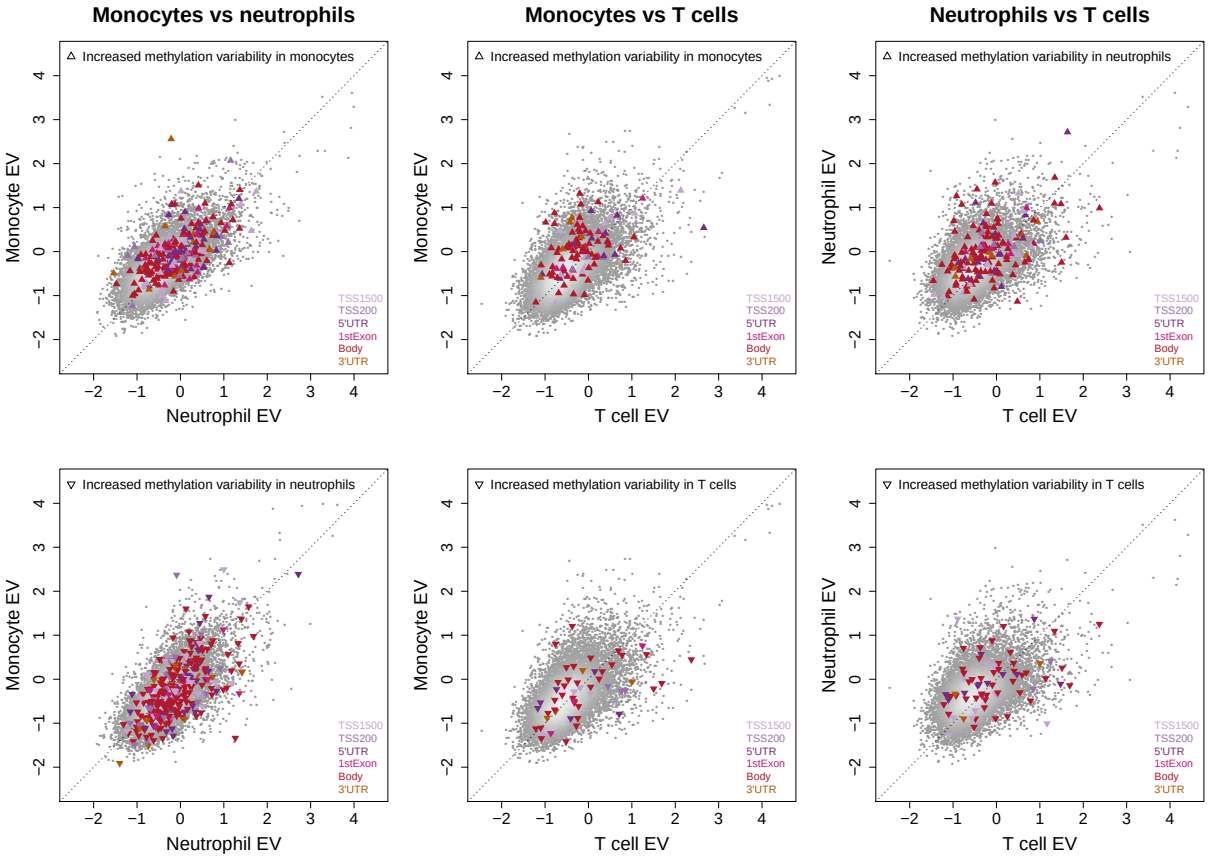


Figure SF26: Scatterplots of EV-scores marking genes with differential DNA methylation variability. Every data point represents a gene. The colors correspond to the genomic region to which the hypervariably methylated CpGs of the genes belong. The dotted line represents the identity line. The three different pair-wise comparisons of the three groups are shown in columns. Top row: Genes with increased DNA methylation variability in the first group are marked. Bottom row: Genes with increased DNA methylation variability in the second group are marked.

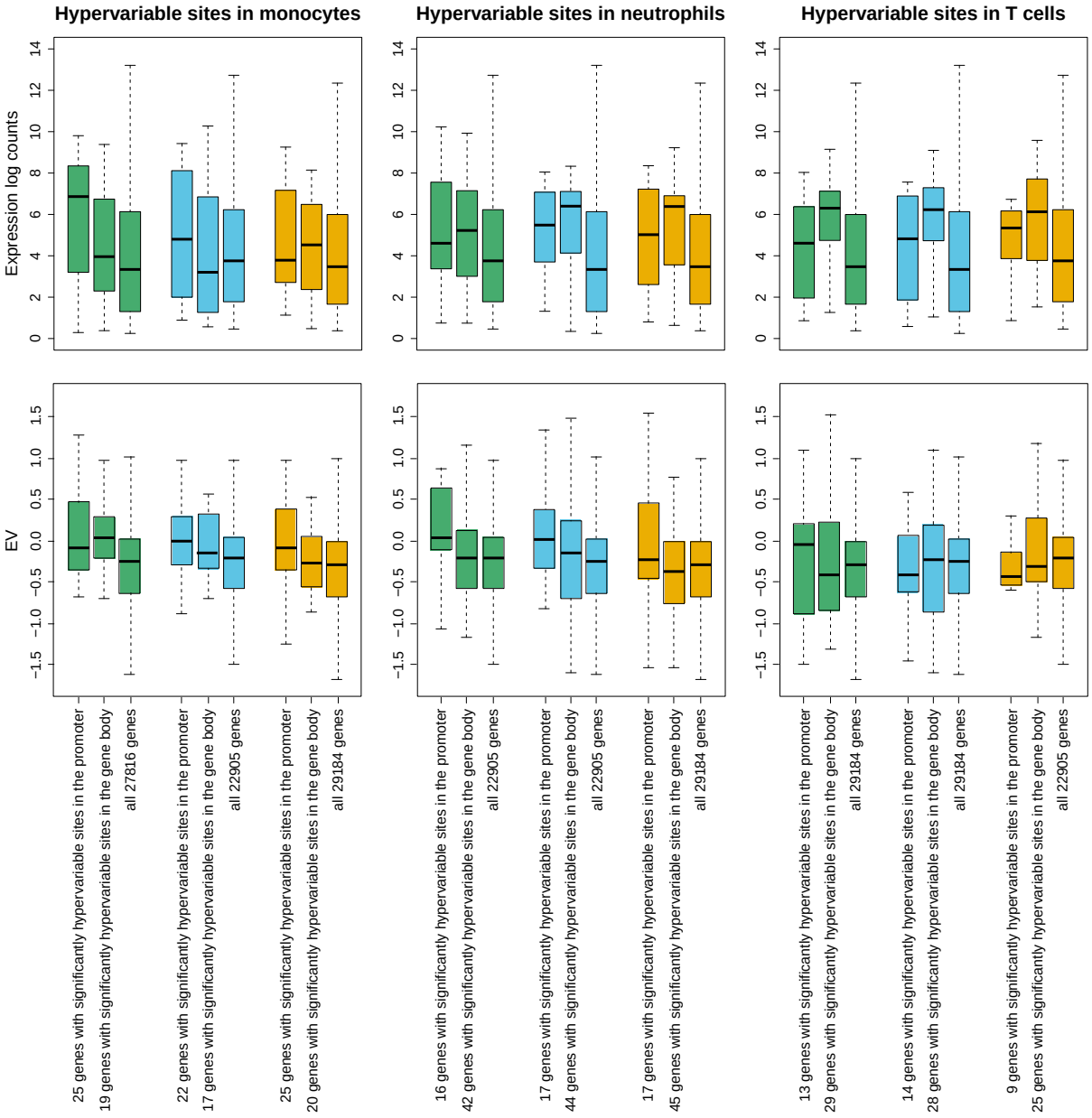


Figure SF27: Boxplots of mean expression levels and EV-scores comparing genes with cell type specific differential DNA methylation variability to others and across cell types. Left column: Genes with cell type specific DNA methylation variability in monocytes in either their promoters or gene bodies are plotted beside the rest of the genes for all three cell types. Middle column: Genes with cell type specific DNA methylation variability in neutrophils in either their promoters or gene bodies are plotted beside the rest of the genes for all three cell types. Right column: Genes with cell type specific DNA methylation variability in T cells in either their promoters or gene bodies are plotted beside the rest of the genes for all three cell types. Top row: Mean expression levels are compared. Bottom row: EV-scores are compared.

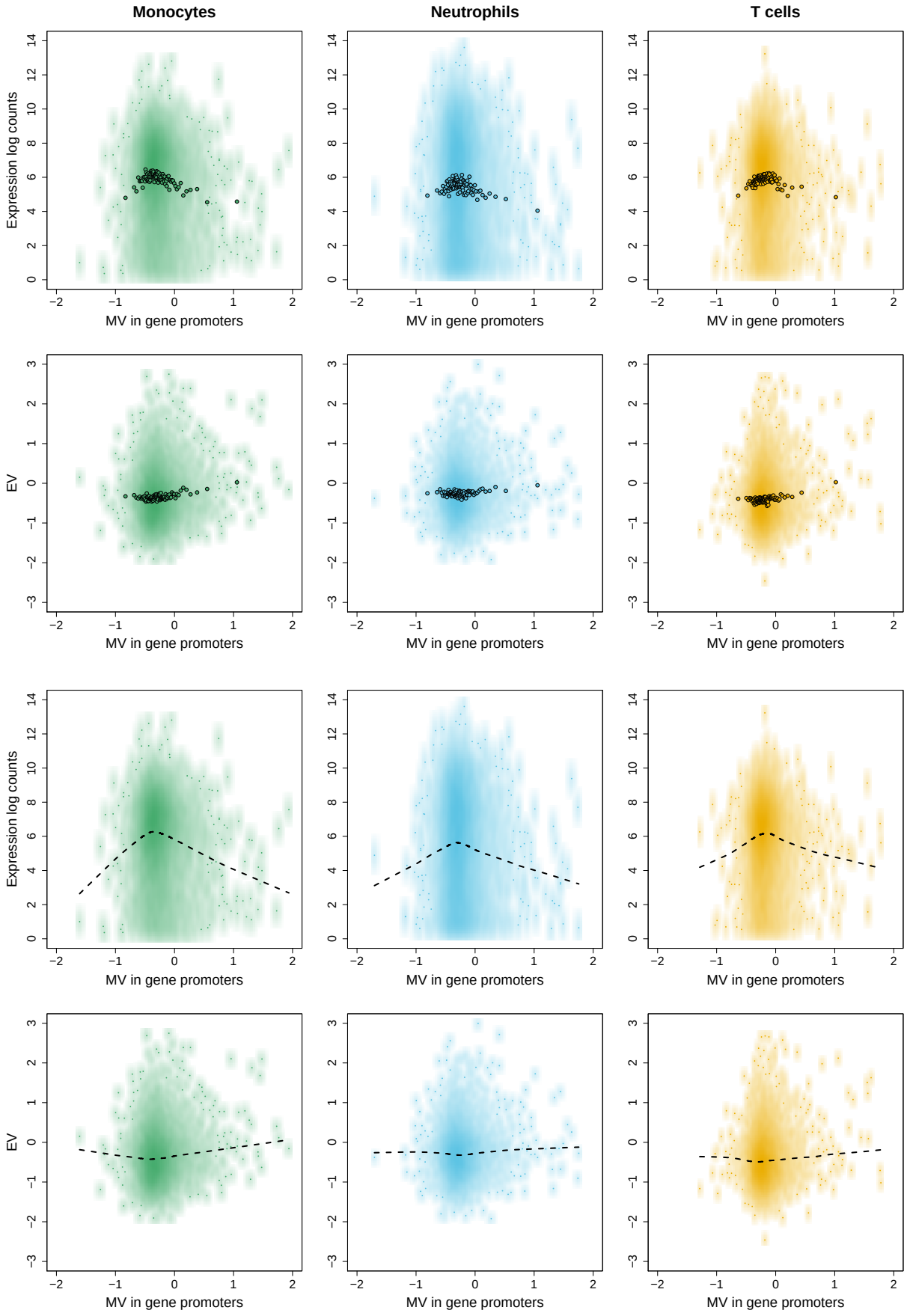


Figure SF28: Global relationship between promoter methylation variability and gene expression. MV-values were plotted against the mean expression values or against the EV. First two rows: Scatterplots of the original values, including the binned data points from figure 4.22 (filled circles). Last two rows: Scatterplots of the original data values, with a lowess regression line.

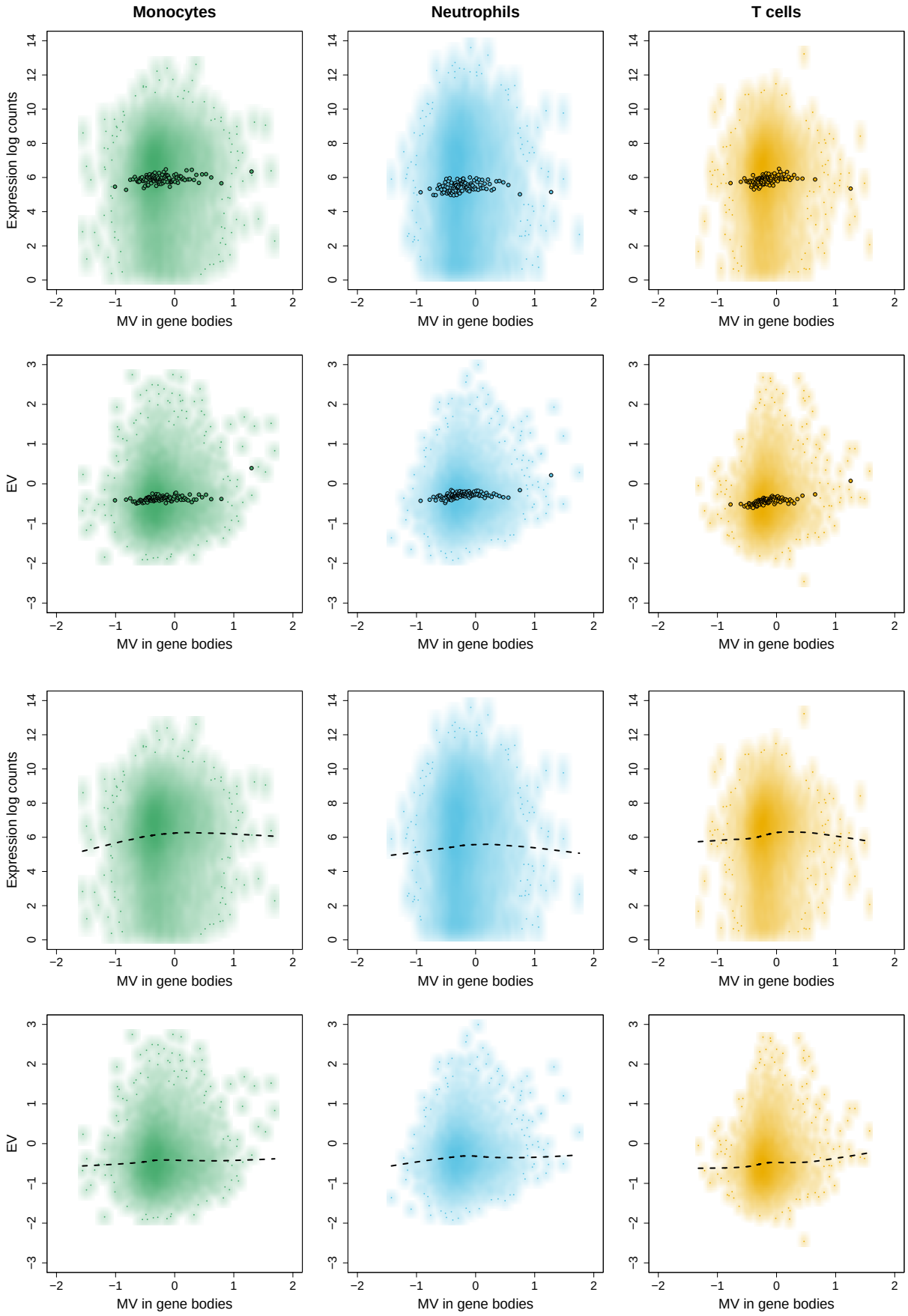


Figure SF29: Global relationship between gene body methylation variability and gene expression. MV-values were plotted against the mean expression values or against the EV. First two rows: Scatterplots of the original values, including the binned data points from figure 4.23 (filled circles). Last two rows: Scatterplots of the original data values, with a lowess regression line.

# Annex II

## Publications

# Publications Forming Part of the Thesis

Kulis M, Heath S, Bibikova M, Queiros AC et al. 2012. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat Genet* 44(11), pp. 1236-42.

Ferreira PG et al. 2014. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res* 24(2), pp. 212-26.

Ecker S, Pancaldi V et al. 2015. Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. *Genome Med* 7(1), p. 8.

Kulis M et al. 2015. Whole-genome fingerprint of the DNA methylome during human B-cell differentiation. *Nat Genet* 47(7), pp. 746-56.